75 NRP
Big Data
National Research Programme

Big Data

Applications, technologies, and societal aspects

Résumé of the National Research Programme “Big Data” (NRP 75)

Swiss National
Science Foundation

Foreword

The ubiquitous collection and processing of data has resulted in profound changes in the way we live. Big data technology is meanwhile vital for competitive, technology-driven economies and holds enormous potentials for scientific advances and for society at large. At the same time, it causes numerous societal challenges. The field as a whole is a top priority for flagship research.

The National Research Programme "Big Data" (NRP 75) targeted fundamental and applied technological research as well as societally oriented research. To this end, the Swiss National Science Foundation launched an open call. An international panel selected 34 research projects from the numerous submitted proposals. NRP 75 did not search for answers to specific technological or societal questions, but instead aimed to advance Swiss capabilities through intensified research in this emerging research area. As a result, Swiss research actors from natural sciences, engineering, and social sciences carried out research projects and initiated public debates on data rights, privacy, and sovereignty, the digital divide, the fairness and accountability of algorithms, and the appropriate level of regulation.

After five years of intense work, the NRP 75 research demonstrates the opportunities for dedicated big data applications and offers a showcase of how high-quality fundamental research made in Switzerland can directly contribute to core technologies. The achievements emphasise that creating solutions requires a global perspective, which includes sound ethical, legal, and socioeconomic dimensions from the start.

NRP 75 complements other research initiatives such as Digital Lives (2018–2019) and the National Research Programme "Digital Transformation" (NRP 77, 2020–2025) that focus on specific societal topics of digitalisation. These actions support the building up of research, innovation and education capacities in Switzerland as envisioned by the Federal Council's Strategy "Digital Switzerland."

The work done within NRP 75 underlines the importance of interdisciplinary efforts to develop and leverage emerging technologies in a responsible way. The challenges include privacy and data security, while allowing access and sharing of data, as well as having the capacities and the talents to imagine, develop and integrate knowledge and expertise in existing processes.

NRP 75 has contributed substantially to reinforcing the necessary competences in big data in Switzerland, the capacities for interdisciplinary innovation as well as the abilities to find appropriate social and legal solutions. This Résumé contains a valuable excerpt of the findings and conclusions of the National Research Programme “Big Data”.

Bert Müller
Delegate of the Programmes Division of the National Research Council
of the Swiss National Science Foundation since January 2021

Friedrich Eisenbrand
Delegate until December 2020

Executive summary

The ongoing societal digitalisation results in the collection of massive volumes of data. This so-called big data holds the potential for widespread societal, industrial, and scientific value creation – if harnessed effectively. The National Research Programme "Big Data"(NRP 75) ran between 2015 and 2022. Its research projects developed real-world applications, invented new technologies and investigated societal aspects in relation to big data, thereby increasing the Swiss research and innovation capacity in big data.

Fifteen research projects produced concrete big data applications in collaborations bringing together computing and domain experts (chapter 2). The results showcase the concrete impact that big data innovations can have in domains such as renewable energy planning, patient monitoring in hospitals, evidence-based policy making, and science itself. The projects underline the importance of interdisciplinary teams capable of working across disciplines and who can navigate the relevant ethical, legal and operational contexts.

Eleven research projects in computer science studied and invented technologies needed for harnessing current and future big data. They covered infrastructure aspects, including data access, cleaning, indexing and pre-processing; and they covered analytics, including query processing, data mining and machine learning, to facilitate knowledge extraction from data (chapter 3). These advances can improve the functionality and performance of big data applications, for instance by enhancing privacy or reducing the computing and data resources needed for model training in machine learning.

Eight research projects investigated specific societal aspects of the deployment of big data, including ethical and legal aspects (chapter 4). Projects studied concrete examples of the use of big data, for instance in human resources and insurance. Their outcomes stress the importance of adapting the legislation to the changing capabilities of technology, of developing guidelines related to the use of big data, of increased ethics awareness and transparency related to the use of data, and of closely following how big data impacts democracy.

Three transversal projects (Cross Cutting Activities) explored the barriers to data sharing in research, supported women's participation in big data science, and produced an overview of societal issues linked to big data.

The fast-paced technological developments in big data continue to offer new opportunities across many domains, including in industrial production, renewable energy, cybersecurity, or e-commerce (chapter 5). These developments also create challenges, in particular an increasing consumption of energy by data processing, the balance between privacy and value creation, the potential risks of discrimination and the need for accountability. Addressing these challenges will be crucial for optimising the value creation from data by companies and public institutions.

The conclusions of NRP 75's Steering Committee suggest how to facilitate responsible value creation from big data, contributing to political and professional debates on this new resource. They are summarised below and are covered in more detail in the report (chapter 6).

Conclusions of the steering committee

Foster an appropriate environment for big data development
(1) Enhance education of big data professionals
(2) Support legal and ethical advice for big data research and development projects
(3) Enable certification of big data application properties

Integrate big data in public and private organisations
(4) Increase the exploitation of big data technologies in the health sector
(5) Strengthen policy making and evaluation with big data
(6) Promote shared data collection, application benchmarks and open-source software

Update and create adequate regulation
(7) Pursue more proactive regulation of big data
(8) Advance data privacy and digital sovereignty in big data applications
(9) Increase transnational harmonisation of regulations

155
CASE

1. Introduction: big data, big changes

Increasingly sophisticated hardware and software technologies enable the collection and analysis of unprecedented volumes of data. These offer the potential for wide-reaching value creation in the public and private sectors. Implementing big data applications in a responsible manner calls for research on all aspects of big data, including computational infrastructures, data analysis methods as well as ethical guidelines and legal frameworks. NRP 75 has made significant contributions across this value chain and has strengthened the ability to invent big data technologies, deploy applications and inform regulation in Switzerland.

1.1 The increasing role of data in society

From digitalisation to big data

Fuelled by public and private scientific research and innovation, advances in technology occur at an accelerating pace and have profound effects on the way we live. Advances in information technology are fostering widespread digitalisation throughout society, yielding a profusion of data that has an ever-greater impact on the way we live and work.

The capacities for data processing, storage and communication have expanded at an exponential rate for several decades. The density of transistors on microchips has been doubling every couple of years, yielding a similar speed up of processors and data transmission.

These spectacular advances are enabled in part by the very large economies of scale in the semiconductor and communication technology industries and by very fast market growth. Coupled with advances in software technologies – in particular operating and data management systems, programming languages and compilers, and analytics methods such as machine learning – they have enabled unprecedented gains in information processing and efficiency, created countless new tools, as well as profoundly changing industries, professional practices and life habits.

Ever more digitalised public, private and personal spheres produce ever larger datasets. Big data commonly refers to datasets whose size and other properties challenge current information and communication technologies, so demanding new solutions. Beyond size (or volume), "big data" also encompasses the speed of data creation and processing, its variety and its veracity (see "When is data 'big'?" p. 11). The scope of big data therefore evolves continuously as technologies advance.

Big data tracks a growing part of our social, professional and individual lives, and of business transactions, industrial devices or scientific research. Some big datasets are personal, some are not. The data is collected by websites, apps, cameras as well as by sensors deployed in smartphones, vehicles, industrial production lines or environmental monitoring systems.

Creating value

Data is regarded as an immensely valuable resource that has yet to be exploited in full. It has been dubbed "the new oil" (or "the new soil") to underline its role as an indispensable resource that fuels many processes and plays a central role in society.

Data has little value in itself. Its value is extracted in the form of actionable outcomes of analytics. Planning is required to identify the questions to be asked and the data that can provide answers. It is necessary to determine how the data can be collected or accessed, to develop efficient analysis tools and to turn the results of that analysis into actions that create value. All along, one must evaluate the possible unintended side effects.

This paradigm is not new. It has been used for decades to support market research, customer surveys, financial planning, or epidemiology. Big data, however, brings a new scale by its sheer size, the required computing and communication infrastructure, the complexity of algorithms needed to analyse it, as well as the scope and challenges of applications. Increasingly diverse examples show how big data can enable more fine-grained models that in turn enable improved business, industry, or government processes.

However, the use of data has generated new risks and challenges for society, in terms of security, fairness, and social cohesion. These will require adequate and proportionate solutions.

Applications in the digital and real worlds

Familiar examples of big data applications include Internet-based services, such as social networks, messaging and e-mail services, searches and advertising, and e-commerce or entertainment services.

The energy and transport domains also use big data. For instance, wind turbines equipped with dozens of sensors can generate data points 20 times a second. This information can be analysed in real time to fine-tune the pitch of the turbine blades and maximise their efficiency. A terabyte of data can be produced by sensors during an aircraft's flight. Properly transferred, stored and analysed, this data helps allow components to be monitored and their maintenance scheduled (ahead of any failures), so optimising fleet management and reducing downtime[1].

Very large datasets are routinely used for environmental monitoring and management, helping for instance the United Nations evaluate loss of biodiversity and follow the effects of climate change[2], or helping predict the distribution of pollutants in air or water[3]. Consortia of journalists have exposed systematic international tax evasion by analysing millions of leaked documents[4].

In research, big data has been produced for decades in international physics collaborations, such as at CERN or by powerful telescopes. It now contributes to a growing number of disciplines, from biology to the humanities. For instance, a major breakthrough in the life sciences occurred in July 2022 with the release a of a huge database of possible 3D shapes of nearly all the 200 million known proteins, as predicted by machine learning algorithms. This is crucial information to understand the roles of proteins in physiological processes, as well as predict their actions and design new treatments[5].

Ensuring responsible use

The power of big data applications also generates numerous risks. Creating value responsibly requires finding appropriate and proportionate solutions to challenges surrounding privacy, ethics, business, law and governance.

For instance, hackers regularly expose private information of millions of individuals. Internet companies have in some cases abused the data of their

When is data "big"?

Big data is an evolving concept as it describes datasets whose properties challenge current technologies.

Their **volume** (size) typically exceeds gigabytes (GB) to reach terabytes (1000 GB) or even petabytes (1000 TB), calling for very powerful storage and processing infrastructure. The **velocity** of the data (the rate of production or transfer, or the analysis speed) can exceed one GB per second, which demands very fast hardware and efficient software.

Applications often combine heterogeneous types of data (text, numbers, coordinates, images, sound, video, etc.) with very different characteristics – a GPS trace being very precise while textual semantics are often ambiguous. This **variety** requires algorithms that can handle multiple data formats and types.

Data is rarely error-free, truthful, accurate, representative or complete – properties captured by the term **veracity.** Many big data applications are based on more or less accurate models or on machine learning techniques that first learn on training datasets of varying quality, which influence the **validity** of results.

Additional V's are sometimes used to describe a big data application, including data variability, vulnerability, visualisation or value.

[1] The case for an industrial big data platform, General Electrics (2017)

[2] World environment situation room, United Nations Environment Programme, https://data.unep.org/

[3] A new area of utilizing industrial Internet of Things in environmental monitoring, HH Lou et al. (2022) Front. Chem. Eng. 4:842514.

[4] The Panama papers: exposing the rogue offshore finance industry, , International Consortium of Investigative Journalists., https://www.icij.org/investigations/panama-papers

[5] The entire protein universe, Ewen Callaway (2022), Nature 608, 15-16. Some 200 million protein structures are stored by the international consortium Uniprot, which grew in 2003 out of the Swiss initiative Swissprot.

La taille du big data
Les applications traitent des jeux de données jusqu’à l’échelle du pétaoctet (un million de gigaoctets).

1 Gigaoctet (Go)

– 1 film

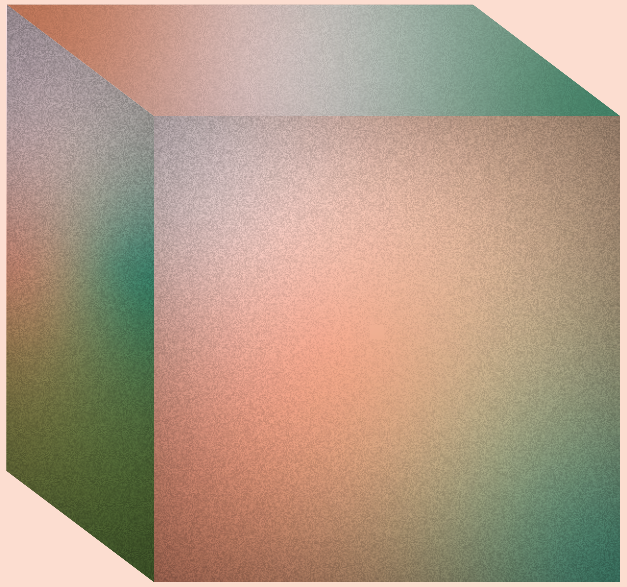
– 1000 livres

1 Téraoctet (To)

– 1 disque dur externe

– 1000 films

– données pour entraîner des modèles de langage artificiels



1 Pétaoctet (Po)

– images médicales produites dans un hôpital en un mois

– vidéos uploadées chaque jour sur YouTube

1 Exaoctet (Eo)

– trafic mobile mondial quotidien

clients, including their medical records, while profiling customers can compromise their privacy. Machine learning algorithms can amplify the bias present in the data used for training and can generate discriminatory results[6]. Platform economy companies, for instance in hospitality and transport, have already disrupted the job market, profoundly challenging work regulations. Addressing such issues is a global endeavour involving the private and public institutions that produce, store, transfer, analyse and use big data, as well as regional, national, and international public administrations, NGOs, and citizens. Rapidly adapting current law and passing new legislation are needed to ensure that applications of big data respect core principles of privacy, fairness, transparency, accountability and non-discrimination.

1.2 The importance of strong research

Scientific research plays a central role in developing the foundations for big data applications. First, it contributes to the necessary infrastructure, such as low-energy sensors producing data for the Internet-of-Things or technologies capable of storing, transferring, and processing vast amounts of data. Second, it provides new tools to produce the analyses and predictions needed by applications, such as advanced statistical methods, data mining algorithms and machine learning techniques.

What's more, research is urgently needed to better understand the societal impacts of big data. Evidence-based recommendations are required to address existing and emerging challenges, including oversight and regulation as well as new practices for business and government. Strong academic research also constitutes a foundation for high-quality education, a central element of the Federal Council's Digital Switzerland Strategy.[7]

Switzerland must have access to state-of-the-art competences in big data

Maintaining highly competitive big data research in Switzerland is of strategic importance for several reasons.

- It is crucial to stay in touch with international developments that shape big data. Swiss research and innovation projects only gain access to world experts and the newest insights if they have something to offer, as the best work only with the best.
- Switzerland's current high-level research and education has attracted important big data companies to the country, and some major corporate research centres.
- Switzerland must make sure it can educate, attract, and retain specialists in big data, which will be ever more sought after internationally.
- Many aspects of big data, particularly the ethical, legal, and societal aspects, are specific to Switzerland so require insights from scientists working in Switzerland.
- Multinational companies often push the limits of big data technologies, and increasingly determine the

[6] Why algorithms can be racist and sexist, Rebecca Heilweil (2020), Vox.
[7] https://digital.swiss

field's course. Keeping some control over technology and direction requires a strong public research community.
- Research helps ensure that the public is informed. Scientists contribute to disseminating new research results and helping the public, the government and private companies make informed contributions to democratic processes and decision-making.
- High-quality research contributes to high-quality education, crucial for a skilled private and public workforce.

These are the key reasons why strengthening big data research is strategically important for Switzerland. An important contribution has been made by the National Research Programme "Big Data".

1.3 The National Research Programme "Big Data"

New insights into infrastructures, applications and societal aspects

The National Research Programme[8] "Big Data" (NRP 75) was designed in 2014/2015. It complements national strategic programmes supporting digitisation, such as the Digital Switzerland Strategy, the cross-industry initiative DigitalSwitzerland, the Swiss Digital Initiative as well as dedicated research initiatives such as Digital Lives.

NRP 75 was allocated CHF 25 million that enabled the funding of a portfolio of research projects that satisfied stringent criteria of scientific excellence[9]. They ran between 2017 and 2021 and belonged to three categories:
- fundamental innovations in the computing infrastructures necessary for big data applications;
- use-inspired research projects developing concrete, real-world applications;
- research on the interplay between big data and society, including legal, ethical and societal aspects.

Four goals were set in the call for proposals of NRP 75:
- achieving advances in computing and information technology;
- addressing societal, economic, regulatory (both local and global), and educational challenges;
- enabling of applications;
- strengthening of research capacities.

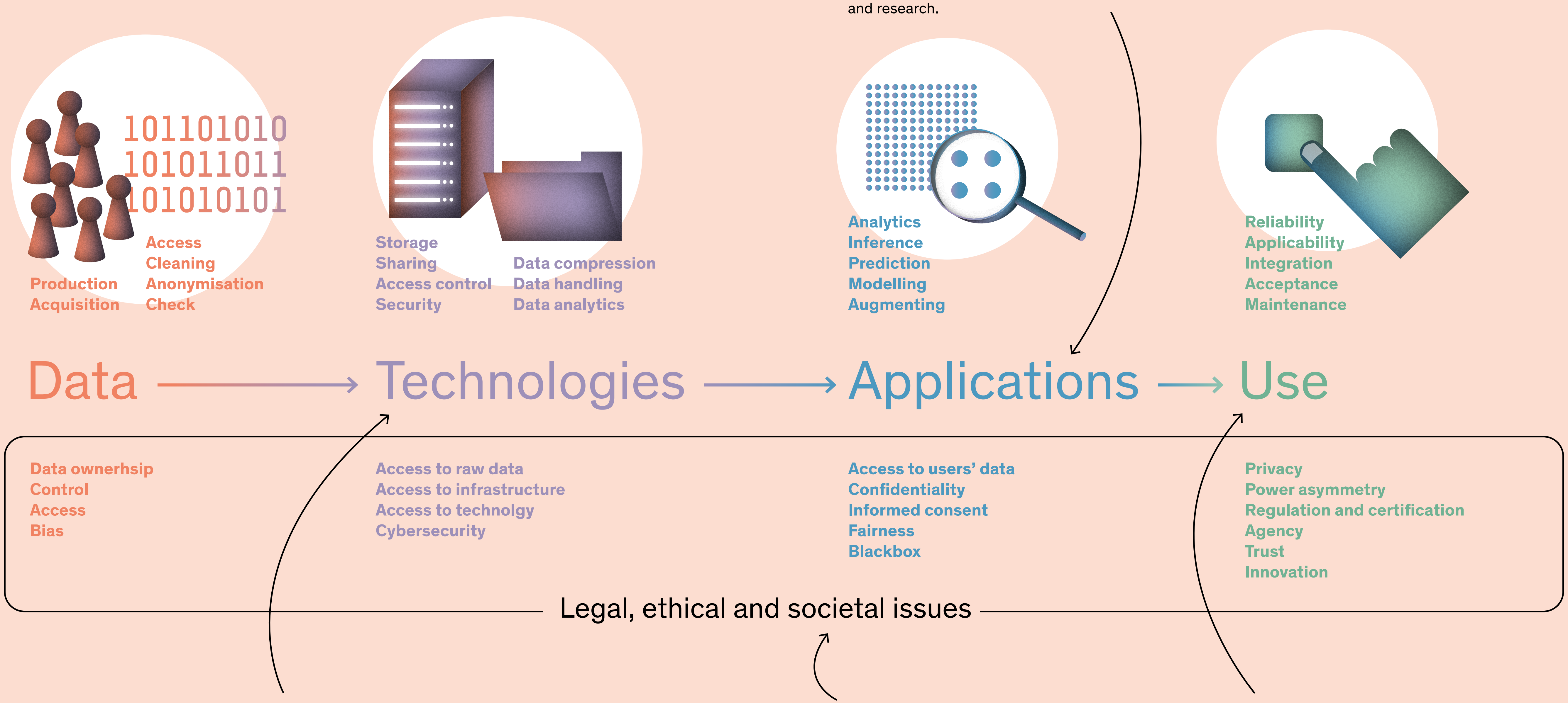
The programme brought scientific advances that contribute to more efficient big data infrastructures, developed concrete applications in several domains, provided insights into ways to address societal aspects, and strengthened the big data research and innovation capacities in Switzerland.

[8] National Research Programmes (NRP) enable thematic research consortia to address topics of importance to Switzerland. They are proposed from the bottom up by administration units, research institutes or individuals to the State Secretariat for Education, Research and Innovation. They are approved by the Federal Council and implemented by the Swiss National Science Foundation (SNSF).
[9] See also Appendix: The National Research Programme "Big Data" (NRP 75) on p. 84.

The big data pipeline

A complex value-chain transforms data into applications. **Data** is produced by online activities or by sensors before it is acquired, checked, cleaned and anonymised. Hardware and software **technologies** are needed for the necessary big data infrastructure to store data, manage access and secure it, and to perform pre-processing and analytics. **Applications** are based on data modelling to produce analyses, predictions or recommendations. These must be validated and integrated into existing processes to ensure concrete **use. Legal, ethical and societal issues** arise all along this value chain, from data privacy to potential algorithmic bias and regulation. The Résumé's chapter cover most of this big data pipeline.

Chapter 2 decribes the applications developed by NRP 75 in domains such as health, sustainability, socioeconomics and research.



Production
Acquisition

Access
Cleaning
Anonymisation
Check

Storage
Sharing
Access control
Security

Data compression
Data handling
Data analytics

Analytics
Inference
Prediction
Modelling
Augmenting

Reliability
Applicability
Integration
Acceptance
Maintenance

Data → Technologies → Applications → Use

Data ownerhsip
Control
Access
Bias

Access to raw data
Access to infrastructure
Access to technolgy
Cybersecurity

Access to users' data
Confidentiality
Informed consent
Fairness
Blackbox

Privacy
Power asymmetry
Regulation and certification
Agency
Trust
Innovation

Legal, ethical and societal issues

Chapter 3 discusses research progress made in big data infrastructures (hardware and software) and in **analytics methods** needed for preprocessing data, for the management of cloud servers or for improving machine learning techniques

Chapter 4 discusses the many societal, legal and ethical issues apprearing in the whole value chain, in particular data ownership, control and access, informed content, power asymmetry and privacy, addressing regulation and guidelines.

Chapter 5 proposes an outlook for big data, discussing the impact, opportunities and challenges brought by its growing use in society.

Looking back: important contributions from NRP 75

NRP 75 was designed in 2014, at a time when many big data technologies and issues that are well known today were only emerging. The very rapid advances made in big data technologies and their deployment in society were a challenge for the research teams, which had to be flexible and adapt their goals. Looking back, the scope of the programme as well as the funded projects have covered essential questions all along the pipeline of big data applications (see "The big data pipeline", p. 16).

The decision taken eight years ago to also include societal challenges in the research programme proved well-founded, as shown by the numerous discussions taking place today about fairness and bias in artificial intelligence, data sovereignty, or the impact of new applications on citizens and employees.

Funding projects to develop concrete applications enabled collaborations to be created across disciplines. These collaborations brought together domain experts, computational scientists, and partners from the public and private sector. This helped strengthen the interdisciplinary knowhow necessary to develop big data applications, creating valuable experience for the decade to come not only in big data but also in data science in general. Such collaborations are expected to play an increasingly important role in addressing global societal challenges as espoused by the UN's sustainable development goals, including climate change, the environmental crisis and aging populations.

Finally, research projects studying the technological challenges of advancing big data infrastructures strengthened the expertise available in Switzerland to develop and shape big data technologies.

Societal impact of NRP 75

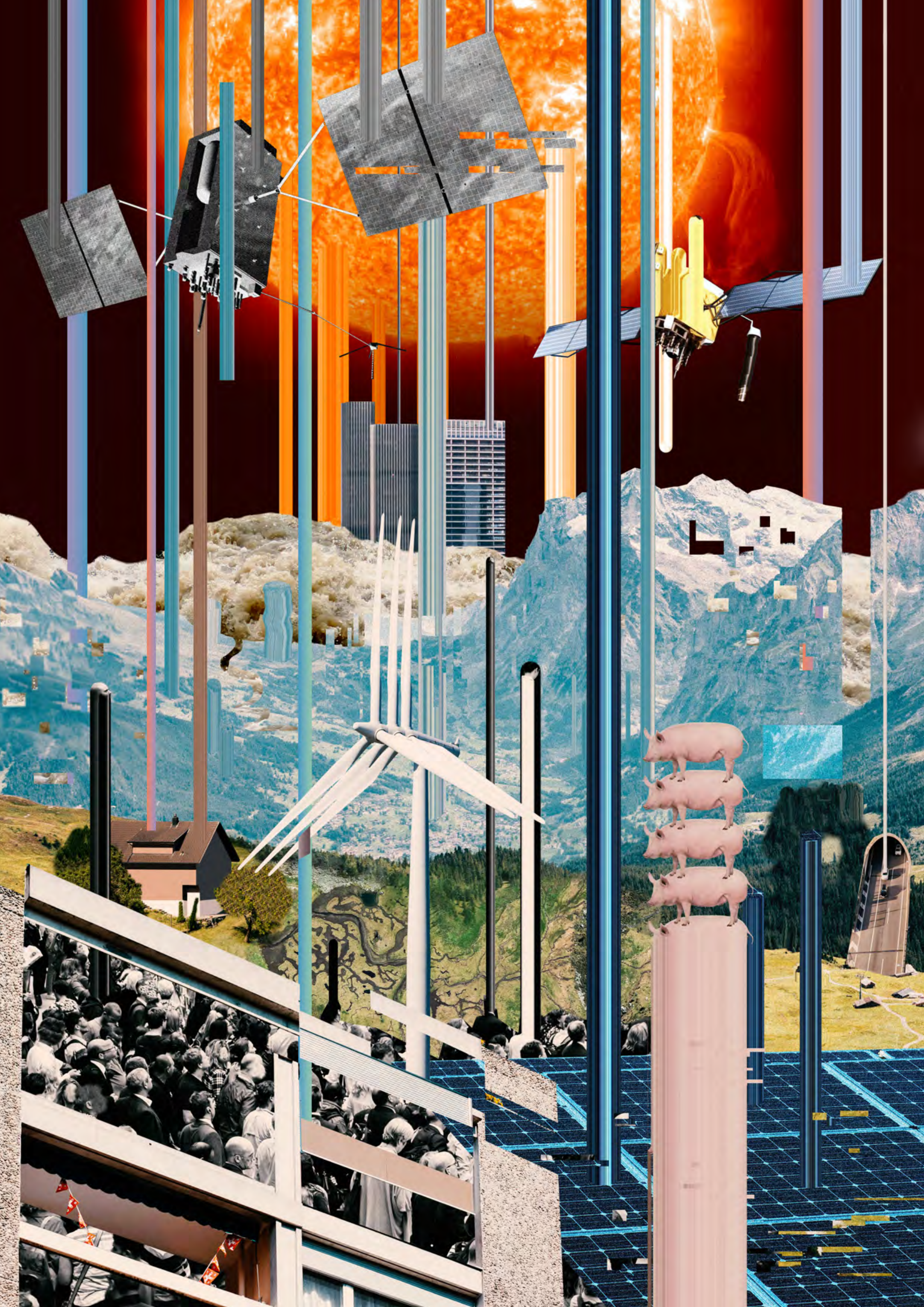
Outreach activities of NRP 75 brought the subject of big data's impact on society to a wider audience. These activities included discussion of legal and ethical aspects of big data, looked at big data and gender equality, and provided schools with teaching material on big data. Overall, NRP 75 helped ensure that Switzerland can benefit from big data in a responsible way. Some projects brought valuable insights for policy making beyond the topic of big data by analysing real-world data, in particular for renewable energy, environmental management and socioeconomics (see "Findings beyond big data", p. 33).

The challenges of big data cannot be solved once and for all, but in order to address them Switzerland must maintain state-of-the art capabilities in research, education, and innovation. NRP 75 focused the attention of Swiss scientists as well as private and public actors on the issues surrounding big data by funding research through an open and competitive call with high quality standards. The programme helped provide a solid foundation to Swiss big data research and to strengthen its relevance and impact. The latter will also greatly benefit from the ongoing National Research Programme "Digital Transformation" (NRP 77) that was launched three years after NRP 75 to focus on the effect of digitisation on education, the labour market, governance and trust.

1.4 The structure of the Programme Résumé

The Résumé of NRP 75 summarises the key findings from the funded research projects and presents an outlook on the challenges of big data.

Chapter 2 “Big data applications” details opportunities brought about by the analysis of large datasets in concrete applications, as explored by NRP 75 projects in domains such as health, sustainability and socioeconomics. Chapter 3 “Big data technologies” summarises advances made in NRP 75 to address technical challenges brought by big data, in particular more efficient computing infrastructures as well as new approaches to data analysis. Chapter 4 “Societal, legal and ethical aspects of big data” offers new insights on the societal challenges of big data, in particular data ownership and privacy as well as fairness. It includes concrete guidelines developed by NRP 75 projects. Chapter 5 “The road ahead” provides a more general outlook for the opportunities and risks related to big data that may become more relevant in the coming years. The conclusions of the NRP 75’s Steering Committee are presented in chapter 6. Finally, the Appendix provides key facts on NRP 75. The Résumé reports on key aspects of big data along the pipeline going from raw data to real-world applications, as shown in “The big data pipeline”, p. 16.



2. Big data applications

Big data applications bring opportunities in numerous domains. Their development requires, however, significant work: forming partnerships with stakeholders, securing technical and legal access to data, developing useful analytical models and validating the applications with users. This chapter presents a dozen new applications developed in NRP 75 projects in domains spanning health, sustainability, policy making and scientific research, and it highlights the lessons learned.

Society's ever growing use of digital technologies generates increasing volumes of data. Many actors would like to exploit this resource to create new tools or make existing ones more accurate and efficient. This leads to a surge in big data applications in numerous domains.

Despite the current hype, building applications based on big data remains a complex and lengthy task fraught with many challenges. The development process requires bringing partners together, solving legal and ethical challenges linked to privacy and fairness, accessing and analysing data, developing models, coding the applications and evaluating their accuracy and usefulness. Only then come the last steps: validating the applications with the users, incorporating them within existing workflows, and ensuring appropriate maintenance.

The National Research Programme "Big Data" (NRP 75) has developed applications based on big data in domains such as health, sustainability, socioeconomics, and scientific research. These represent only a small subset of the sectors that are exploring or integrating big data applications, which range from traditional data-intensive domains such as banking, marketing, and health, to new fields such as agriculture, journalism or policy making.

The projects funded by the programme improved existing methodologies and developed new ones for domain-specific big data applications and highlighted the potential benefits to society and the economy, such as the development of strategies for personalised medicine, smarter transport planning, integrated deployment of renewable energy and clearer evaluation of the effect of socioeconomic policies.

The various NRP 75 applications have reached different stages of development, including initial models, prototypes and fully-fledged systems. This variety echoes the opportunities and challenges of creating practical applications for big data more generally.

The applications developed within NRP 75 and their the key messages from the projects are grouped into four domains:
— healthcare in section 2.1;
— sustainability, including transport, energy, food supply and environmental management, in section 2.2;
— socioeconomics in section 2.3;
— and research in section 2.4.

Section 2.5 presents a summary of NRP 75's insights into building big data applications. The research projects are described in section 2.6.

2.1 Improving and personalising healthcare

Numerous new approaches are being pursued to tailor healthcare to the specific characteristics and needs of individuals and population groups, following what is known as the "P4" paradigm of predictive, preventive, personalised and participatory medicine. This approach relies on better access to data, developing robust analytical tools, and fostering close collaborations with practitioners and patients to match the new digital tools to their needs. Health data originates from traditional databases such as electronic health records (EHRs) and from new sources such as smartphones and wearable sensors – technology that is becoming ever more commonplace.

As medical records are considered to be sensitive information, big data applications must comply with stringent data protection laws. These laws vary across jurisdictions, posing a challenge to cross-institutional collaborations, as discussed in chapter 4. Applications of big data can potentially have a significant impact on health research, education and care, reaching beyond medical institutions into people's homes.

The NRP 75 research projects on healthcare (see their description in section 2.6) constitute only a very small subset of the possible applications. However, they provide concrete examples of how healthcare can benefit from big data:

— allowing for more efficient management of critical situations in intensive care units by monitoring patients and devising accurate predictions of their status (see the NRP 75 research project *Intensive care units*);
— prototyping a smartphone-based system to personalise the management of lower back pain (see project *Back pain management*);
— developing new techniques for biomedical research (see projects *Big genetic data, Genomes comparison* and *Bioinformatics databases,* which will be discussed also in section 2.4.)

Key messages on healthcare applications

The NRP 75 research projects demonstrate the potential for big data applications to support modern and efficient healthcare, but also reveal many challenges. In particular, developing useful and practical applications requires robust infrastructure for sharing data, including appropriate solutions for technical, management and legal issues.

Societal impact

Automating the analysis of patient monitoring can improve the quality of care by providing alerts to staff that have a low rate of false alarms *(Intensive care units).* Mobile apps can help patients carry out physiotherapy exercises at home and also gather data that can help evaluate the impact of therapy *(Back pain management).* Biomedical research can also benefit from big data (see section 2.4): several methodological and technical advances are improving the science of genomics in terms of clinical research, epidemiology and environmental biology *(Big genetic data, Genome comparison),* as well as helping a wider range of specialists exploit biomedical databases *(Bioinformatics databases).*

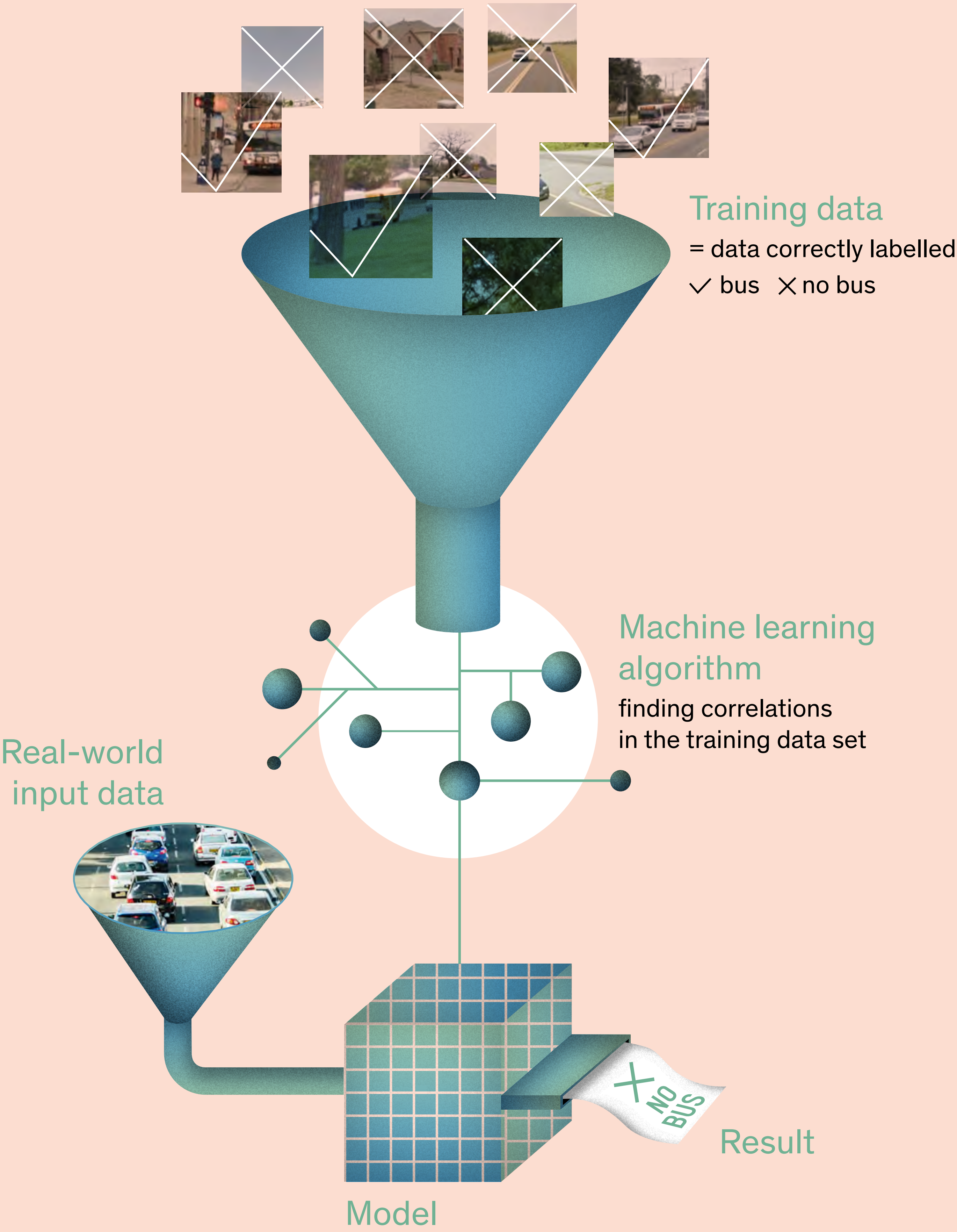
Impact on big data capacities in Swiss healthcare

One application succeeded in integrating multimodal data (video, sensor data, medical records) in a machine learning algorithm *(Intensive care units).* Another combined implicit and explicit information, showing that data collected by sensors can be used to assess the veracity of observations such as patients' self-reports on their exercising habits *(Back pain management).*

Challenges

It is necessary to develop new ways of collecting and managing healthcare data, which is generally kept in segregated data silos. In particular, putting in place a national healthcare system based on more transparent and widely shared data would facilitate new digital healthcare solutions *(Back pain management).* Bringing together various data sources is hindered by the high heterogeneity of the Swiss health systems and regulations, and by the fact that the data collected by stand-alone medical devices is highly proprietary (owing to a lack of commercial and regulatory

How supervised machine learning works
An algorithm learns to solve a task (here: recognising a bus) from labelled data. The larger the training dataset, the more accurate the results are on real-world input.



Training data
= data correctly labelled
✓ bus ✕ no bus

Machine learning algorithm
finding correlations in the training data set

Real-world input data

Result

Model

incentives to integrate the data) *(Intensive care units).* The prospective collection of labelled health data and managing regulatory processes for data handling take considerable time and effort.

Mobile health apps offering personalised physical exercises for rehabilitation or pain management can help patients adhere to treatment protocols but are ultimately still dependent on patients' motivation *(Back pain management).* Digitising healthcare data can also lead to public mistrust of government and private health companies, as exemplified by criticism of the contact tracing application SwissCovid App.

Solution avenues

A number of measures can be taken to improve the development of healthcare applications, and favour the use of big data in healthcare.

— Generalise data collection for particular diseases across medical centres in Switzerland, as the traditional set-up of individual study centres linked to specific hospitals or Cantons is not suitable for big data health applications. This is necessary to generate enough data to provide insights into complex disease. Attention should be paid to the advances made by multinational technology companies in gathering, storing and analysing private and public health data.

— Streamline the processes necessary for legal compliance, such as using digital signatures or obtaining patients' consent for the use of their data in medical research. Ensure that regulatory bodies such as ethics committees are well-informed about the latest developments in big data applications *(Back pain management);* see also chapter 4.

— Underline the need to harmonise data formats used by medical devices and promote greater use of standard formats in Electronic Health Records (EHRs).

— Tell all stakeholders that it is important to maintain high-quality metadata, which is crucial to efficiently indexing and searching large databases *(Intensive care units).*

2.2 Supporting sustainability

Building sustainable societies requires optimising the interactions between the numerous components of energy, transport, supply or food systems. For instance, patterns of mobility depend on both the supply of transport and the demand from citizens (who have varying needs and preferences), but also on complex interactions determined by factors such as the weather, acceptance of home working and the availability of financial incentives.

Big data applications can significantly improve systems' efficiency and sustainability. This often requires analysing fine-grained data in real time in order to monitor individual activity such as the movement of people on public transport or the performance of solar panels on specific roofs. Such insights help continuously optimise supply to meet expected demand, by adapting transport timetables or varying electricity production, for example.

NRP 75 presented half a dozen innovative applications with a clear potential to support sustainability:

— generating high-resolution maps at regional and national scales showing how solar, wind and geothermal energy could benefit individual

buildings *(Renewable energy potential);*
- creating high-resolution 3D digital models of cities *(City digital twins)* and using GPS traces to explore how collective and private transport could be optimised *(Optimising transport management);*
- creating tools to automatically quantify the extent of eroded terrain in aerial images *(Soil erosion)* and using mobile or surveillance videos to automate flood detection *(Flood detection);*
- developing a pilot platform that collects and analyses data relevant to the Swiss pig industry *(Pig data).*

Key messages on sustainability applications

Sustainability can strongly benefit from big data applications, given the long track record of using data in engineering, and for efficiency and security monitoring. Ongoing digitalisation, including deploying devices for the Internet of Things, generates ever bigger datasets, thereby improving prospects for big data applications. These applications require data from different sources to be integrated and powerful new analytical tools to be developed. Society's urgent need for sustainability calls for greater leverage of big data.

Societal impact

The applications developed within NRP 75 show that the use of big data can help support sustainability. In particular, it can aid evaluation and adaptation of national and local strategies on renewable energy, create realistic scenarios for transport management, produce 3D maps for urban planning, monitor environmental degradation and support the management of natural hazards.

Impact on big data capacities in sustainability

NRP 75 has shown that new techniques can augment sparse datasets to create models of energy production at high resolution and with quantified uncertainties *(Renewable energy potential).* New algorithms can create high-resolution 3D numerical models of urban areas thanks to vehicle-mounted cameras scanning continuously in all directions *(City digital twins).* It is also possible to gain insight into single mobility traces from global, aggregated and anonymous databases *(Optimising transport management).* Supervised machine learning can identify the main factors impacting soil erosion and landslides *(Soil erosion).* Efficient interfaces for visualising and interacting with information at different scales can be designed by integrating end-users early in the design of applications *(Flood detection).*

Challenges

As with many other applications, those designed for sustainability are dependent on the availability of comprehensive and high-quality data. The environmental data needed to deploy renewable energy devices is often lacking or incomplete *(Renewable energy potential),* while access to data on people's movements needs support from industry, particularly communication providers. Current privacy regulations can hinder such access, but there are ways to limit the risk of data leakage such as by providing access only to a secure storage backend *(Optimising transport management).*

Supervised machine learning algorithms that are trained with labelled data do not always work reliably in real-world applications. These can sometimes benefit from more advanced approaches, such as unsupervised learning, which does not use training data. Important metadata, such as videos' GPS location, are often absent from crowdsourced data, which makes analyses much more difficult and time-consuming *(Flood detection).* Natural variations in aerial photography – shade, mowed fields, etc. – make it difficult to analyse how terrains change in time *(Soil erosion).*

It is hard to develop big data applications in industries where data use remains fragmented and where analyses are hindered by data's low quality, limited scope or lack of consistent ontologies (the way information is conceptualised, categorised and represented). This is also the case in industries where confidentiality is important and data is shared reluctantly *(Pig data).* Finally, staff changes in partner organisations may bring a planned collaboration to a sudden halt.

Solution avenues

- Develop privacy-by-design approaches that integrate solutions to privacy issues from the beginning of the project and involve experts on legal and societal aspects early on *(Optimising transport management, City digital twins);* see also chapter 4.
- Develop policies making it possible and easier to share sensitive data between industry and academia for research purposes while safeguarding confidentiality and privacy *(Pig data).*
- Foster a culture of digitisation. Increase awareness of the issues and challenges of data usage and support continuous education in this domain. Include staff with experience in technical and applied sciences to bridge the gaps between academics and industry, especially when data culture is still emerging *(Pig data).* Establish trusted governmental platforms for sharing crowdsourced data of interest to the public and ensure they include the necessary metadata *(Flood detection).*
- Involve end-users early on to understand their needs and develop practical tools for interpreting and visualising data *(Flood detection).*
- Invest in unsupervised learning and other approaches that reduce dependence on potentially skewed training data *(Flood detection).* Conversely develop supervised machine learning when adequate training data is available *(Soil erosion).*

2.3 Better understanding of socioeconomic interactions

The use of sophisticated data analysis approaches is increasingly being explored for social and political applications. However, the necessary data is not always readily available. When it is, analysis can be more complicated than in the natural and technical sciences: causes and effects are usually hard to unravel from data because they are multifaceted and embedded in differing socio-political settings. The fact that more data is being collected nevertheless offers many opportunities for new applications to help evaluate, inform and possibly improve policies.

Concrete applications in socioeconomics have been developed within the NRP 75 to:

- improve methods to find causal relations in socioeconomic datasets, in order to inform and improve policy making *(Evidence-based policy);*
- scale up analytic methods to study patent production and citations,

mapping innovation and knowledge generation worldwide (*Mapping global innovation*).

Key messages on socioeconomic applications

Societal impact

The increasing availability of socioeconomic data offers great opportunities for better evidence-based decision and policy making. However, analyses are often harder than people think. The project *Evidence-based policy* has developed techniques based on machine learning to untangle causal effects from mere correlations. Big data techniques can also map the spread of ideas and influence worldwide, which should help better understand what leads to successful innovation *(Mapping global innovation).*

Impact on big data capacities in socioeconomics

Using algorithms to analyse politically relevant data can help reduce bias when researchers unconsciously search for results fitting their own views. This approach can also establish whether certain subgroups of people react differently to policy measures compared to the average population, allowing for finer-grained insights and policies. Such techniques are now ready to be applied in practice *(Evidence-based policy).* New methods can now be used to analyse large networks and can contend with temporal effects, helping identify the nodes having rapid or lasting impact *(Mapping global innovation).*

Challenges

Even careful data analyses can be biased by the methods they use. Relationships may be hidden because data is too granular *(Evidence-based policy),* and datasets have intrinsic limitations – for instance, the fact that patent citations cannot be followed across different national registers *(Mapping global innovation).*

Solution avenues

- Promote the use of advanced analytics based on socioeconomic data in policy evaluation *(Evidence-based policy).*
- Encourage data aggregation at different granularities, such as at the population and sub-population levels, and promote causal analysis to infer sound relationships from data *(Evidence-based policy).*
- Foster international compatibility of national datasets *(Mapping global innovation).*

2.4 Accelerating research

Big data is also supporting innovation by playing an increasing role in fundamental research. This is seen particularly in very large international

Big data in and for society 1

Supporting big data literacy in schools

Learning to use information and computing technologies is part of the Swiss obligatory school programmes (Medien und Informatik, Education numérique, Tecnologie e media), as specified in the three coordinated curricula. This involves acquiring data skills, which requires new and optimised teaching supports. In collaboration with the Museum for Communication, NRP 75 produced teaching material on big data for lower and upper secondary school levels, comprising eight modules of two lessons each. Launched in early 2020, the material has been downloaded more than 12 000 times. The initiative was promoted by the Museum of Communication in Bern and the Swiss Academy of Engineering Sciences (SATW).

Big Data: Lehrmittel für die Sekundarstufen / Big Data: outil pédagogique pour les cycles secondaires, NRP 75 and Museum of Communication Bern (2020) collaborations such as the Human Genome Project or the experiments at the Cern particle physics laboratory in Geneva as well as in the high-throughput observational and analytical systems used for example in astronomy, satellite sensing or genomics. A growing number of scientific disciplines have established standards for data generation and sharing. The increasing size of datasets – which can be as large as several petabytes – makes it harder to access, store, manipulate and analyse the data in question. This prompts the need for efficient analytical methods and machine learning algorithms.

NRP 75 produced new methods and tools to improve the handling and analysis of massive research databases in chemistry, solar physics and genetics:

- accelerating the discovery of new molecules thanks to efficient simulation of their properties *(Computational chemistry);*
- improving the understanding and prediction of solar flares, which can generate geomagnetic storms on Earth *(Solar eruptions);*
- facilitating the analysis of genomic and biological data in biomedical research, supporting the development of new therapeutic and diagnostic approaches *(Big genetic data, Genomes comparison, Bioinformatics databases).*

Key messages on research applications

Societal impact

Developing the right analytical tools, especially in machine learning, helps to exploit increasingly large research datasets. This enables faster and broader discovery as well as more accurate prediction in life sciences, chemistry and space physics.

Impact on big data capacities in research

It is possible to achieve high prediction accuracy even with very large datasets that have been compressed to reduce the time and cost of data transfer, storage and processing *(Solar eruptions, Big genetic data).* Machine learning techniques can dramatically reduce the time needed to accurately solve very complex mathematical problems, such as identifying the properties of molecules *(Computational chemistry).* Several NRP 75 projects developed tools to manage and search massive biological datasets at the petabyte (million gigabyte) scale *(Big genetic data, Bioinformatics databases).*

Challenges

Labelled data is often unavailable in research projects, limiting the use of standard supervised machine learning tools *(Solar Eruptions).* At the same time, metadata –information about the data itself – is often of low quality, preventing efficient data access, indexing, and analysis *(Big genetic data).*

Solution avenues

- Encourage academia and industry to share and curate large research datasets within the paradigm of Open Research Data (see “The challenges to Open Research Data”, p. 45).
- Use real datasets to evaluate developed methods early on, so that the research scope can be adapted if needed.
- Invest in world-class fundamental and applied research projects that use large datasets to foster big data skills in academic and private environments.

2.5 Key messages on big data applications

Beware the hype

Reporting by the media, industry and think-tanks can give the impression that big data is a magic tool: just find the data, add some machine learning, train the algorithms, build an application, and you are ready to disrupt professional practices and entire industries. This overly simplistic vision hides numerous hurdles of a conceptual, technical, legal management and collaborative nature.

Building a big data application requires immense effort in many steps: creating partnerships with stakeholders; finding the data; assessing data quality, preparedness and completeness; storing the data securely, preparing the data for analysis; finding suitable existing algorithms and then adapting them or creating new ones; benchmarking the results; creating practical visualisations and interfaces to explore the results; and, finally, integrating the new application into established workflows.

Over-celebration of big data obfuscates very deep, if seemingly obvious, questions. Does the data exist at all? Is it accessible? Is it properly described with metadata? Can privacy be preserved and regulations respected? Do end-users actually need the intended application? Such questions must be considered at the outset in order to realistically gauge the amount of work ahead.

Is your domain big data ready?

Big data applications can be developed in a fairly linear way when they build upon existing prototypes *(Renewable energy potential, Optimising transport management)* or when high-quality and standardised data exist, as in weather, cartography and biological domains *(Solar eruptions, Big genetic data).* This allows computer scientists to focus mainly on technical issues, such as establishing a pipeline to access data in real time, building algorithms, or creating user-friendly interactive interfaces. NRP 75 projects made a number of important advances in the design, implementation and evaluation of practical approaches to data engineering, including data management, analytics, visualisation, evaluation, auditing, integration and mining.

Conversely, it is much harder to build big data applications in domains that are less digitised, lack a data culture, and are averse or unused to sharing data *(Pig data, Mapping global innovation).* In such cases, significant time and effort is needed to deal with non-technical issues, such as setting up partnerships between stakeholders unused to sharing data. The availability and quality of data needs to be assessed early on, potentially leading to a revision of the application's scope.

Big data demands interdisciplinarity

Managing datasets at the petabyte scale requires much time, manpower and collaborative effort to solve numerous technical and sociolegal challenges.

Building an application that makes an impact usually requires an interdisciplinary approach in which the various stakeholders address all potential issues and questions early on, including how the foreseen solution will be used by end-users *(Pig data, Flood detection).* Applied scientists, domain experts, industry partners, research engineers, and ordinary citizens need to interact frequently to ensure that high-quality data are acquired and shared, as well as ensuring that applications meet the needs of their intended users. Scientists from NRP 75 have explored new ways of interacting effectively with different stakeholders. Involving users early on improves the design of applications *(Flood detection).* This kind of experience ensures that academic research is able – when needed – to rapidly develop real-world applications.

The real world is messier than training data

Many machine learning algorithms can fail in the real world, with potentially grave consequences when it comes to health prediction or autonomous vehicles. This is a problem for supervised learning when algorithms are fed incomplete, noisy or unrealistically uniform training data. Turning to more frugal unsupervised learning could yield more robust systems *(Flood detection).*

Think privacy-by-design from the start

Issues related to privacy and regulation need to be addressed carefully, in particular with the help of legal experts (see also chapter 4). Privacy-by-design approaches should be considered and implemented as early as possible. This requires the careful consideration of overarching principles such as purpose limitation, transparency and proportionality, as well as data minimisation, accuracy and security *(Intensive care units, City digital twins).* Sharing experiences, within and across domains, helps establish best practice.

2.6 The research projects on big data applications

NRP 75 developed applications based on big data in a variety of domains: two in healthcare, six in sustainability, two in socioeconomics and five in research capacities.

Research projects in healthcare applications

Intensive care units: an automated alert system

This project developed an “ICU-cockpit” to help staff in neurosurgery intensive care units react quickly to critical situations. These involve in particular ischemic (brain) strokes and epileptic seizures. The platform could help medical staff prioritise interventions and increase patients’ safety. It involves a collaboration between the Neurocritical Care Unit at University Hospital Zurich, ETH Zurich and IBM Research Zurich.

The system predicts critical situations by integrating and analysing many types of data, including electroencephalography, video streams and patients’ medical history including brain imaging and laboratory analyses. The team created technologies to capture biomedical data in real time at a high resolution up to 200 Hertz. They designed and implemented algorithms that could automatically detect epileptic seizures based on video and electroenceph-alography, and predict imminent secondary brain injuries. The application was based on the data from more than 100 patients with a type of stroke known as subarachnoid haemorrhage. Two algorithms were developed to reduce the rate of false alarms, one using machine learning and the other video monitoring of patients' motion. The system was integrated and tested in a clinical environment.

—

ICU-cockpit: IT platform for multimodal patient monitoring and therapy support in intensive care and emergency medicine
Emanuela Keller (University Hospital Zurich)

Back pain management: a personalized smartphone-based solution

This project developed a smartphone-based approach for managing back pain. The team built the app "Swiss Health Challenge" to collect, transmit and store anonymised sensor data. In addition, they evaluated three different preventive strategies to reduce the high costs of treatment (which usually involves painkillers, physiotherapy, and surgery). Researchers from ETH Zurich developed machine learning methods to analyse the data within a collaboration involving the University Hospital Balgrist and the Swiss medical device company Hocoma.

One preventive approach regularly asked patients suffering from low back pain to perform interactive exercise sessions at home. While a physiotherapist's regular assessment did not reveal improvements amongst non-severe patients, the study offered new insights into patients' adherence to physical exercise programs and on how fear of movement can affect postural sway (the small unconscious movements that maintain balance).

The app was complemented by movement sensors to identify physical stress during ski training, which can lead to back injury. This approach provides far more detailed information about exercise behaviour than self-reports. The project also highlighted the importance of patients' motivation when adhering to a physical exercise programme.

—

Personalized management of low back pain with mHealth: big data opportunities, challenges and solutions
Robert Riener (ETH Zurich), Walter Karlen (Ulm University, formerly ETH Zurich)

Research projects in sustainability

Renewable energy potential: evaluation for Switzerland

This project created a digital platform to evaluate the potential of geothermal, wind and solar photovoltaic energy for the heating and cooling of buildings. The resulting nationwide estimates have a high resolution in space and time, and can help plan energy systems at local and regional levels, as well as optimise incentives and revisit the national energy strategy.

The system integrates data on the weather (wind and solar radiation), the environment (topography, geology and ground temperatures), and the built environment (roof orientation and the space available for boreholes). It uses these data to create regional and national maps of the renewable energy potential, with a spatial resolution on the scale of individual buildings and a temporal resolution of about an hour. New mining techniques for big data were able to interpolate the available measurement points to fill gaps within the maps. For instance, wind maps with a scale of 250 meters were generated for the whole country based only on data produced by 208 MeteoSwiss monitoring stations. The team also

Findings beyond big data

Several NRP 75 research projects produced results that have direct societal relevance, notably in the domains of sustainability and socioeconomics. They illustrate the potential of big data for supporting policy.

The project *Renewable energy potential* produced concrete findings that can **support the national energy strategy 2050.**

→ It is possible to reach 50 % of the nationwide potential for solar photovoltaic electricity with only 10 % of existing rooftops, by focusing on those with the highest potential. This corresponds to around 12 TWh of electricity per year, or around 20 % of national consumption.

→ Around 1 000 wind turbines could generate 4 TWh, the goal for wind power in 2050 set by the Federal Office of Energy.

→ High-resolution maps of the shallow geothermal potential in Vaud and Geneva indicate a potential production of 4 TWh of heat, or 40 % of the current demand in the two cantons. Using district heating networks to distribute the heat amongst settlements would double this heating potential.

→ Reinjecting heat into the ground during summer is important for sustainable geothermal energy use. In combination with district heating networks, shallow geothermal energy can cover more than 70 % of the heating and cooling demand of Swiss buildings by 2050.

The project *Soil erosion* produced new insights into **erosion in Alpine regions.**

→ From 2007 to 2016, an increase of 80 % of the area subjected to soil erosion was observed in a 2000 km² region centred around Martigny (VS), corresponding to 11 % of the Swiss Alps.

The project *Evidence-based policy* established causal effects in **socioeconomics topics.**

→ Training people to search for jobs does not, on average, enable jobseekers to find employment more quickly. However, it can increase the chances of subgroups such as migrants of finding a job by up to 60 %.

→ Practising music has a positive effect on cognitive and non-cognitive development in children

→ Football referees are more likely to penalise teams coming from specific linguistic areas in Switzerland.

The project *Mapping global innovation* produced **insights into innovation** by analysing millions of patents.

→ Patent ecosystems are less international than expected. Patent citations often cluster geographically and in terms of disciplines: more interdisciplinary patents do not appear to be more successful. Patents from some countries receive particularly high numbers of citations – Switzerland, for instance, being known as a prolific hub of innovation.

→ Companies and organisations play different roles in the way patents cite each other, some of them acting as central hubs and knowledge distributors.

See the project descriptions for more details.

quantified the generated maps' uncertainty. The project produced numerous results relevant for energy policy (see "Findings beyond big data", p. 33).

—

Hybrid renewable energy potential for the built environment using big data: forecasting and uncertainty estimation
Jean-Louis Scartezzini (EPFL)

Optimising transport management: anonymous individual mobility traces

This project explored ways to gather and analyse smartphone GPS data to gain insights about people's mobility patterns. More than 4000 participants installed a dedicated app, which allowed the recording of more than a million trips. The scientists involved developed new processes to anonymise the data, identify the kind of transport involved (whether walking, cycling, or using a bus, car, etc.) as well as the activity (sport, work, education, etc.). The project's results were used in other large mobility studies in Switzerland, such as ones examining the accuracy of self-reports (Swiss Mobility and Transport Microcensus) or the response to mobility pricing (Mobility Behaviour in Switzerland). The results were also used to generate and evaluate different scenarios, such as changing public transport timetables or instituting new traffic management, in a mobility simulation platform called Matsim. They showed for example that during the Covid-19 pandemic there was a huge decrease in the use of public transport in Zurich and a significant increase in cycling.

The project was carried out with telecom provider Swisscom. It uncovered the potential of GSM traces for transport modelling and optimising related policies. It showed that individual – but anonymous – mobility patterns can be extracted from aggregated GSM traces, avoiding the need to actively collect data from individuals and compromise privacy.

—

Big data transport models: the example of road pricing
Kay W. Axhausen (ETH Zurich)

City digital twins: 3D models from a scanning car

This project produced a car-mounted camera system to scan urban areas. It designed a camera that records in all directions and an algorithm to produce a complete 3D numerical model of a city from these continuous scans. Such virtual "digital twins" of urban areas can support city planning and transport strategies.

The project also studied legal and ethical issues, and developed a privacy-by-design approach based on the principles of purpose limitation, transparency, proportionality, as well as data minimisation, accuracy and security. For instance, it automatically removed sensitive details, such as car number plates and faces when it was tested in Sion (VS). The team has been discussing technology transfer with a Swiss startup offering urban digitisation services.

—

ScanVan – a distributed 3D digitalization platform for cities
Frédéric Kaplan (EPFL)

Soil erosion: quantification by aerial photography in Switzerland

This project created machine learning algorithms able to automatically identify and map eroded soil in official aerial photography. It studied ten sites, mainly in mountainous regions, and showed for instance that eroded surfaces in the Urseren Valley between Realp and Hospental (UR) almost tripled in 16 years to reach 0.4 square kilometre. It identified the main factors influencing erosion and landslides – namely a terrain's slope, roughness and orientation. The automated analysis was used to carry out a more detailed study of a 2000 square-kilometre region centred around Martigny (VS) that covered a tenth of the Swiss Alps. It estimated that the degraded area had risen by 80% between 2007 and 2016. These results make erosion visible to policy makers and can feed into measures used to protect soil within agriculture, tourism and land use planning (see "Findings beyond big data", p. 33). This is a crucial endeavour as soil is a non-renewable resource essential for food production, biodiversity, and management of natural hazards. A follow-up research project funded by the Federal Office of Environment will develop a tool to map erosion on large scales.

—

WeObserve: integrating citizen observers and high throughput sensing devices for big data collection, integration, and analysis
Volker Roth (University of Basel)

Flood detection: automatic geotagging of crowdsourced videos

This project produced the first building blocks for a platform to assist in the management of flood risk. It comprises automated video analysis to recognise crisis situations, establishing a video's location, and presenting the results to crisis managers.

The project made clear the need for

new unsupervised machine-learning algorithms. The existing technology for image recognition, based on supervised learning trained with annotated datasets, can quickly fail when analysing real-world videos.

In a study carried out in the canton of Basel-Land the team tested how to automatically locate videos when taken, for example, with a smartphone lacking GPS information. It did so by comparing the videos with existing street-level images, but this strategy relied on several factors: sufficient high-quality sources, the right weather, and the presence of recognisable infrastructures or landmarks. Working together with experts who manage natural hazards, the scientists tested prototypes of interactive maps to visualise flood scenes. The project passed several milestones in the use of AI for automatically detecting crisis situations.

—

EVAC – Employing video analytics for crisis management

Susanne Bleisch (FHNW)

Pig Data: analytics for the Swiss swine industry

This project brought together stakeholders of Switzerland's swine industry to integrate information on a variety of subjects, such as pig transport, billing, health and treatment, meat quality and fattening processes. It developed methods for real-time predictions. For instance, it confirmed that most farmers provide carcasses that meet prescribed standards of quality while identifying areas in which farms fall short of industry specifications.

The project provided answers to 6 of the 18 questions put by stakeholders, such as what influences the quality of carcasses, how the swine industry network is structured, or what is the difference in quality and revenue between single-breed herds and mixed batches. It highlighted the importance of digitising the breeding industry to key stakeholders such as the Federal Food Safety and Veterinary Office, thereby helping to establish a Swiss centre of competence and information on pig health.

The project revealed how hard it is to carry out analytics in an industry with many small players unaccustomed to working with data– as compared with other fields or to the swine industry in other countries. Specific challenges include the quality and compatibility of data as well as a reluctance to share information.

—

Pig data: health analytics for the Swiss swine industry

John Berezowski (University of Bern)

Research projects on socioeconomics

Evidence-based policy: uncovering causality from data

This project has helped to identify cause-effect relationships in large socioeconomic datasets, developing new machine learning techniques to demonstrate causality as opposed to mere correlations. Specifically, it uncovered cause-effect relationships within real datasets regarding unemployment benefits, education and sports (see "Findings beyond big data", p. 33).

The new methodologies can uncover subgroups that benefit from interventions even when the population as a whole remains unaffected – findings that would remain hidden within averaged data and which could potentially gear policy towards more personalised measures. This project made significant advances in methodology, such as comparing and evaluating existing statistical approaches.

Socioeconomic datasets are becoming increasingly complex, which could mean greater insights but also more difficult analyses. The project developed machine learning approaches to automatically answer one key question of modelling: which factors should be analysed explicitly and which should instead be considered as confounding? Such questions are central to the development of evidence-based policy making and will be explored in a follow-up project funded by the National Research Programme "Digital Transformation" (NRP 77).

—

Causal analysis with big data
Michael Lechner (University of St.Gallen)

Mapping global innovation: analysing patents

This project analysed millions of patents to uncover how innovation is spreading across the globe. It merged several databases containing information on patents and companies worldwide, and created a model to capture temporal changes and effects in complex networks – so helping to identify influential patents and with that insights into global innovation (see "Findings beyond big data", p. 33). The team adapted a statistical model for big data networks with over a million nodes and made it available as open-source software.

—

The global structure of knowledge networks: data, models and empirical results
Alessandro Lomi (Università della Svizzera italiana)

Research projects on research capacities

Computational chemistry: discovering new molecules

This project improved artificial intelligence methods to simulate molecules, a stepping-stone to accelerating the discovery of useful compounds in health and industry. It developed novel machine-learning methods to compute the main properties of molecules, using existing databases containing information on hundreds of billions of molecules. It is very hard to analyse such huge databases efficiently, but the team produced training and testing datasets and was able to characterise molecules much more quickly and accurately than with existing methods, even for larger compounds.

—

Big data for computational chemistry: unified machine learning and sparse grid combination technique for quantum based molecular design
Helmut Harbrecht (University of Basel)

Solar eruptions: predicting geomagnetic storms

This project demonstrated how solar flares can be predicted ahead of time, helping to prepare for potential geomagnetic storms on Earth. These storms can disrupt critical infrastructures such as telecommunications, electrical grids, satellite operations and flight routes. The team used 30 terabytes of solar observations, developed methods to analyse the data in a compressed form, and showed that signals at ultraviolet wavelengths, almost unused at present, can help predict when solar flares will form. The algorithm foresaw a solar eruption half an hour before its occurrence. The team developed both supervised algorithms (with training data) and unsupervised techniques to overcome the problem of limited labelled data.

—

Machine learning based analytics for big data in astronomy
Svyatoslav Voloshynovskiy (University of Geneva)

Big genetic data: powerful indexing

This project created a new method to index very large databases of genetic sequences while compressing them a thousand-fold. This public tool, Metagraph, can help users worldwide efficiently search repositories of genetic sequences and carry out comprehensive analyses. The approach has numerous applications in personalised medicine or for tracking variants of pathogens. It helps tackle the ever-increasing size of genetic databases such as The Cancer Genome Atlas. These databases often contain petabytes (millions of gigabytes) of data and their use requires huge amounts of engineering and work on infrastructure. The project indexed over 1.4 million whole genome sequences. An interactive platform contains the genetic information on 500 000 plants, 450 000 bacteria and 120 000 fungi, as well as on 240 000 human gut metagenomes.

—

Scalable genome graph data structures for metagenomics and genome annotation
Gunnar Rätsch (ETH Zurich)

Genome comparison: faster analysis

This project developed new machine learning methods to compare the genomes of different organisms despite variations in data quality. Such comparisons improve understanding of the evolution of groups of genes involved in specific metabolic processes, revealing which are associated with essential housekeeping functions as opposed to evolution. In particular, the project developed methodologies to find genes (or proteins) with the same function in humans and in model organisms such as the fly or the mouse. This type of research is crucial to large-scale efforts such as the European Reference Genome Atlas consortium or the Earth BioGenome Project, which aim to sequence the genomes of all known animals, plants and fungi. The project's approach could potentially speed up genome comparisons significantly.

—

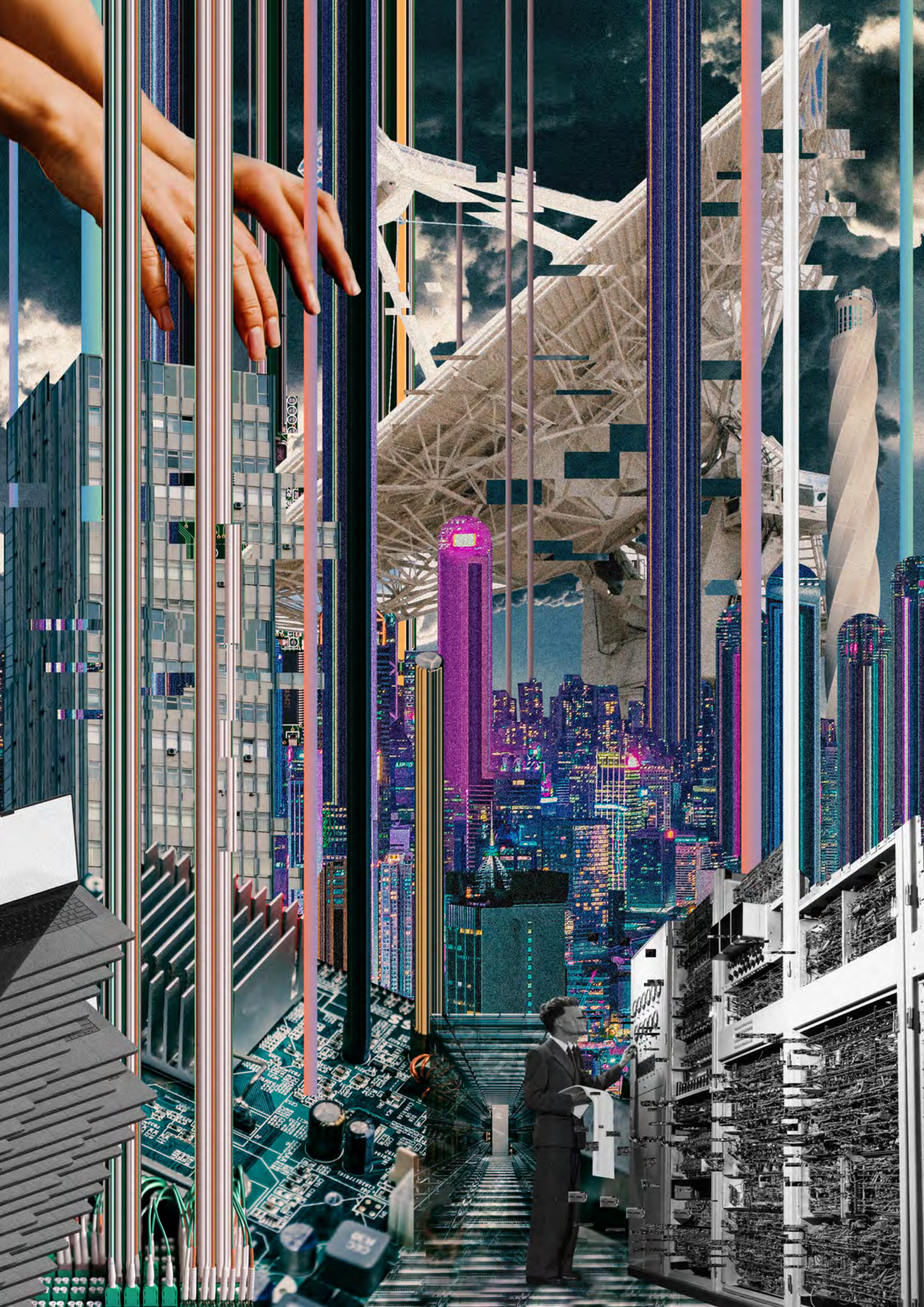
Efficient and accurate comparative genomics to make sense of high-volume low-quality data in biology
Nicolas Salamin (University of Lausanne)

Bioinformatics databases: queries in natural language

This project designed a user-friendly interface to query bioinformatics databases using natural language. The platform provides visual guidance to help keyword searches during the exploratory phase and incorporates feedback to refine queries and improve results. The system integrated databases of proteins (UniProt), gene expression (Bgee) and gene expression across species (OMA). The approach could also be used beyond the life sciences and is being followed in the European project INODE encompassing cancer research, astrophysics and policy making for innovation.

—

BIO-SODA: enabling complex, semantic queries to bioinformatics databases through intuitive searching over data
Kurt Stockinger (ZHAW)



3. Big data technologies

Big data applications require big data technologies: hardware and software solutions able to handle massive datasets as well as analyse them reliably and efficiently. The research outcomes of NRP 75 presented in this chapter show how Swiss academic experts can contribute significantly to developing new technologies for big data applications and can help successfully deploy next-generation solutions for the necessary infrastructures and analytics.

The application of big data in a real-world setting faces profound technological challenges. One is the sheer volume of the data: by popular definition, big data exceeds the capabilities of existing data capture, storage, management and analytics technologies. Many current computing infrastructures will soon be outdated and need replacement. Big data calls for new means of information processing and data analysis.

This is why fundamental research in big data infrastructure and analytics technologies is crucial. It lays the foundations upon which future applications will be built, and also ensures the continued supply of highly skilled experts in the field. Switzerland produces world-class research in computer science and must continue to support it. It should also support new infrastructure as well as the production and curation of reliable datasets – so preparing it for the technological challenges of big data.

NRP75 has strengthened Swiss basic research in this field. It has produced a dozen new approaches to developing the technology that underlies big data applications. These approaches can be grouped in two domains:

- the IT infrastructure, mainly software, needed to access and clean, store, index, control, pre-process and monitor data (see section 3.1);
- data analytics, namely algorithms to extract knowledge from data (see section 3.2).

Section 3.3 describes the challenges of research on big data technologies. Section 3.4 summarises the key messages, while section 3.5 presents the individual projects.

3.1 More efficient big data infrastructures

Big data requires high-performance infrastructure, namely the low-level processes that serve as the backbone for higher-level data analytics. This infrastructure comprises hardware and software.

- Hardware includes processors such as CPUs, GPUs and TPUs (central, graphical and tensor processing units), transient memory and permanent storage, communication elements, powering and cooling elements, etc.
- Software infrastructure enables data access and pre-processing, programming, monitoring of data streams or hardware, etc.

Improving infrastructure for big data therefore requires advances in both hardware and software. While industry drives progress in hardware, academic research contributes significantly to new software that enables faster and more efficient processing of large datasets.

NRP75 explored, proposed and tested half a dozen approaches to improving infrastructure, mainly software:

- contributing to the emerging field of in-network computing, namely the processing of data during its transfer, performed at communication nodes *(In-network computing);*
- improving analytics of dynamical networks arising from streaming data *(Graph analytics and mining);*
- automatically analysing the descriptions of heterogeneous data in order to combine and prepare them for

Key technology concepts in big data

Metadata Information about a data point such as where and when it was acquired, its type or categorisation.
Anonymisation Removing all data points that could reveal a person's identity, so making data anonymous – or rather "pseudonymous".
Re-identification Combining several anonymised datasets to identify people.
Artificial intelligence Algorithms and machines demonstrating "intelligent" behaviour, as well as the underlying methods and real applications.
Machine learning Computing techniques allowing algorithms to learn autonomously, such as via training data.
Supervised learning A machine learning approach in which algorithms learn from labelled training data.
Unsupervised learning A machine learning approach in which algorithms uncover features within datasets without using labelled training data.

further processing *(Loosely structured data);*
- monitoring in real-time data compliance with prespecified rules *(Data streams);*
- creating improved interfaces to handle large datasets in the Scala programming language *(Scala programming language).*

The programme covered only a small subset of the many questions raised by big data technologies and of the approaches being currently explored worldwide.

Impact on big data infrastructures

Processing data within communication nodes while it is being transferred can significantly increase processing speeds and reduce latency for operations on big data *(In-network computing).* New parallelised algorithms can monitor in real time whether a fast stream of large data complies with specified rules. Methods based on formal logic, while often considered much slower than machine learning approaches, can be scaled up *(Data streams).*

Novel computing architecture can improve analysis of massive dynamical networks *(Graph analytics and mining).* A tool for automatically classifying different types of data can help develop applications that combine numbers, text, images or videos *(Loosely structured data).*

The release of a new version of the Scala programming language in 2021, which makes it easier to interface with large datasets, will enable new applications of big data in many sectors worldwide. Invented in Switzerland, Scala is used by major technology, banking and media companies *(Scala programming language).*

3.2 Novel approaches to big data analytics

Analytics is the most visible component of big data applications, creating value from the data by extracting knowledge and insights that are valuable for users or customers. Modern algorithms for data analytics, data mining and multi-dimensional analytical queries combine advanced statistical methods and machine learning algorithms such as deep learning. The challenge for such algorithms is to analyse huge datasets reliably, accurately and in a reasonable amount of time.

In spite of the processing power provided by distributed computing and dedicated processors, it may still take days for algorithms to learn from training data. Speeding up this step requires either more efficient algorithms or clever ways to prune or compress the input datasets, providing good results with less training data.

NRP 75 improved big data analytics in six different ways, including developing efficient approximations to speed up data processing as well as reducing the amount of training data without unduly affecting accuracy.
They
- improved the continuous analysis of very large datasets while preserving privacy *(Stream analytics);*
- reduced the size of the dataset used to train machine learning models *(Coresets);*
- reduced the time needed to train algorithms used for modelling data

and making predictions *(Fast prediction algorithms);*
- allowed efficient real-time monitoring of computing farms *(Data centres);*
- extended our understanding of the limits of deep neural networks *(Machine learning models);*
- improved language models for automated conversational agents *(Language models).*

Impact on big data analytics

NRP 75 research made progress on several types of big data analytics. Novel techniques allow quicker training of algorithms to analyse so-called Gaussian processes, used to model data and make predictions, while quantifying their uncertainty more precisely *(Fast prediction algorithms).* Some data processing such as image analysis can now be performed in real time on an incoming data stream, without having to wait for all of the data to be stored. This idea can also be used to continuously safeguard privacy while data is generated and transmitted: adding random noise to the data obscures individual records *(Stream analytics).*

Better understanding the limitations of deep neural networks, such as whether they are generalisable and can be used in different contexts, helps to keep them robust and reliable *(Machine learning models).* Theoretical advances in the field of conversational agents, meanwhile, boosts agents' sophistication and reliability *(Language models).*
Data can be reduced without compromising reliability. One can significantly reduce the data used to train machine learning models, thereby also reducing the required computing resources *(Coresets).* In data centres, even limited data on computing resources can be used to monitor and optimise processes *(Data centres).* A new method to synthesise tabular data minimises the risk that proprietary or confidential information is leaked *(Data Centres).*

3.3 Challenges in big data technologies research

Challenges

Research projects designed to improve big data infrastructures and analytics face some of the same challenges. The data needed to develop new infrastructure software can be hard to get hold of because companies are not always willing – or able – to provide large and realistic case studies. This can be due to business confidentiality and privacy regulations such as GDPR *(Data centres),* as well as staff rotation, changes in internal policies, or unrealistic expectations of big data applications *(Data streams).* As such, some projects had to resort to synthetic and benchmark data, which is not always suitable for real-world applications. Overall, the procedures regulating data access, processing and sharing, in particular those regarding privacy and consent, are perceived as rigid, heavy and time-consuming *(Graph analytics and mining).*

Research on big data is very competitive, particularly given the involvement of multinational tech companies. These organisations invest heavily in R&D, have unrivalled computing infrastructure and access to the world's largest datasets, and have also been recruiting world-class experts. Academia can collaborate with private industry at a high level, but also needs to find niches where it can compete with the much better funded and equipped industrial teams *(Language models).* Long-term funding and flexible research plans are needed to seize opportunities in this fast-moving field.

Pursuing state-of-the-art research means hiring the best scientists, but the competition from multinational tech companies makes this difficult. Also, academia attaches little weight to innovation, such as developing prototypes with public and private partners or creating and sharing open-source software. This provides little incentive to pursue such projects *(Scala programming language).*

Solution avenues

Current administrative processes regarding data access, sharing and processing in Switzerland can be streamlined, especially when these relate to public research. The use of data in applications can affect privacy adversely, but restricting data use in a blanket fashion has drawbacks, for instance making innovation prohibitively time consuming. Privacy has a cost that must also be considered.

Solution avenues comprise making privacy an inherent and possibly mandatory aspect of big data processing. Developers and users of big data applications must be informed about the various techniques that preserve privacy and their pros and cons. Ideally, they would have access to tools that could help them to optimise algorithms given the balance they want to strike between privacy, efficiency and quality of services. Digitisation of information requires careful analysis to ensure purposeful use of formats and metadata.

Metrics of scientific success for promotion and funding allocation should go beyond the usual publication and citation counts, to potentially include the impact of research outside academia (especially when the work is using open-source protocols). Scientists must have freedom and flexibility of research so that they can tailor their plans to make the most of rapidly evolving fields such as big data.

The human factor can be as important as access to technology, especially since the latter is often open-source and available. Supporting academic research not only enables advances in big data technology locally and internationally, it also trains the specialists that society needs. These experts will not only develop technologies but also bring understanding of the issues surrounding big data, such as technology availability, privacy, cybersecurity and participation of stakeholders. As such, they will contribute to public and private organisations' strategic decisions on digitisation.

3.4 Key messages on big data technologies

NRP75 produced numerous world-class research results, pursuing novel avenues to improve the infrastructure and analytics needed to exploit big data. While such fundamental research is intrinsically very challenging, there is a known route to success and it is ultimately more straightforward than developing applications. For example, limited access to data can sometimes be circumvented by using artificially generated datasets – whose known properties allow testing and tuning of the new systems. The research remains dedicated to the question of how fast the systems can process and analyse data and reach the expected result within a given margin of error. In other words, the research problems are well defined. However, they are embedded in a fast-moving environment with actors in industry and elsewhere pursing different goals.

The private-public research competition and cooperation

The intense international competition in big data technologies threatens nations' digital autonomy, but also provides an opportunity for collaboration. Industry, particularly in the US and China, is making many of the advances in infrastructure and analytics for big data – presenting a good third of the work at top scientific conferences. It leads in the development of language models, image generation and hardware such as Tensor Processing Units, which are optimised to run neural networks. Private research and development is at least on a par with the best academic research worldwide.

Companies' big data technology might appear universal, given their desire for wide adoption. A CPU or pre-processing algorithm may be essentially agnostic as to how they are used. But big data technologies are becoming increasingly specialised to best deal with the problem at hand, and in particular the kind of data involved – whether dynamic or static, homogeneous or heterogeneous, etc. This means that industry is also influencing the possible range of big data applications. It is therefore crucial for publicly-funded research to keep pace with industry if society is to have a voice in the future of digitisation.

Academic research remains essential for developing big data technologies, especially when addressing objectives that are important for society but less so for big tech companies, such as reducing energy consumption or ensuring privacy-by-design. In addition, public research can be bolder by pursuing high risk-high gain avenues. While industry often follows one-size-fits-all approaches, academic research has successfully developed a wider range of hardware and software for big data technology. These include programmable network switches, in-network analytics, new programming models for domain-specific devices, and algorithms based on formal logic instead of machine learning.

Academic research is often far ahead of private industry, the latter relying on the innovation of university spin-offs. But this gap is much smaller for some research topics in big data, so encouraging collaboration between academia and industry – particularly since the former needs the computing power, storage capacities and data access of the latter. Such collaborations are in principle "win-win", with academia gaining from industry resources, real-world problems and a stiff challenge, while industry benefits from state-of-the-art research and more innovative ideas.

One dormant issue in academia is the lack of recognition given to researchers who develop applications, stimulate collaboration, and adopt open-source software. This can potentially deter world class public researchers from addressing concrete problems and collaborating with industry. This calls for more diverse career paths in public research, and metrics that go beyond traditional scientific publishing and funding successes.

The challenges to Open Research Data

Sharing results and data can make research cheaper and more efficient. Data can be reused and combined, while new studies can learn from the body of scientific results. Sharing increases productivity because it reduces the time needed to collect data, while stimulating creativity by enabling low-cost and innovative experiments. It also broadens access to research results, including to people outside academia. In reality, however, research data is often inaccessible and underused. Scientists wanting to share their data still face many practical, institutional and financial hurdles.

NRP 75 addressed these issues with the cross-cutting activity big data: *Big data: open data and legal strings*. It looked at the concrete challenges facing researchers when sharing, publishing and reusing data. It conducted interviews, formulated clear recommendations for research institutions and proposed concrete advice in the form of a How-to Guide.

The main challenges faced by researchers

→ Practical and legal hurdles hinder access to existing research data: low data quality, outdated formats, identification of sources, diverse storage locations, inaccessible repositories, outdated websites, data's legal status and restrictions on re-use, data protection laws, infringement of others' rights, etc.

→ Researchers also face organisational, financial and legal hurdles in publishing research data: lack of incentives, costs, technical and legal know-how, long-term commitment, risk of being scooped by other scientists, risk of data misuse by external partners.

→ Lack of standardisation of research data limits the potential for reuse.

Solution avenues

→ Legal advice: data ownership, intellectual property and copyright law, contractual agreements with third parties, and data protection law.

→ Standardisation: data formats, storage, anonymisation processes, regulatory procedures.

→ Identification and publication of best practices and of a How-To-Guide.

→ Financial and reputational incentives for sharing.

Big data: open data and legal strings
Sabine Gless (University of Basel)

The staffing issue

A major challenge in deploying big data is the scarcity of qualified personnel all along the value chain, from infrastructure technologies to applications, business integration and regulation. There is fierce competition for talent, with many of the brightest minds being hired by large multinationals, as well as mid-size and startup companies. Academic research is losing out as a result, struggling to attract the best scientists – even at PhD level. Universities therefore risk losing talented researchers when they collaborate with big tech companies. Rapid and frequent career changes, while bringing new perspectives and connections, are a problem for research projects.

On the other hand, the world-leading Swiss research on big data ensures that the many specialists needed by public and private organisations are educated and trained, and maintain good contacts with academia and industry. This makes Switzerland innovative and attractive to multinational companies and international organisations.

Getting the data

The second big challenge is the availability of large, high-quality datasets, which are essential for realistic evaluation of big data analytics and applications (see chapter 2).

This problem will diminish as public and private organisations develop a data culture, but its resolution will require a sound strategy that ensures data are high quality, properly described with metadata, and protected by privacy-by-design practices. Developers and users of big data technologies must be familiar with the various techniques for preserving privacy and able to find the right balance between privacy, efficiency and quality of services.

Privacy and data protection raise many questions, such as whether the Swiss or European data protection regulation sets the appropriate boundaries in data management or how to protect privacy while encouraging innovation. The current ethical, approval and administrative processes framing the use of medical and scientific data in Switzerland are perceived as slow and complex, and could be streamlined and simplified. But this is a multi-faceted discussion which requires multi-disciplinary approaches, including the involvement of the social sciences. It is covered in chapter 4.

3.5 The research projects on big data technologies

NRP 75 explored a dozen diversified approaches to improve the technologies needed by big data, half of them in the field of infrastructure, half in the field of analytics.

Big data infrastructure

In-network computing: solutions for graph analytics

This project made several advances in the analysis of large graphs (networks) using in-network computing, namely the processing of data while in transit and before storage. For example, phone calls and social media posts build highly complex graphs, whose size and availability are rapidly increasing.
The new methods rely in particular on network devices such as new generation field programmable gate arrays (FPGAs) or programmable application-specific integrated circuits (ASICs). The results improve on the performance of widely used software-based systems by several orders of magnitude, demonstrating in-network processing of over four billion events per second.

—

Exploratory visual analytics for interaction graphs
Robert Soulé (Università della Svizzera italiana)

Graph analytics and mining

This project extracted inferences from a network and explored graph analytics on different platforms – including combinations of in-core and out-of-core processing. Some results have also proved valuable for storage systems used in general big data processing.

—

Building flexible large-graph analytics and mining systems on commodity hardware
Willy Zwaenepoel (University of Sydney, formerly EPFL)

Loosely structured data: new tools for integration

This project developed tools to combine different types of data – such as text files, pdfs or images – and prepare them for processing and analysis. This issue of data variety is a tricky one, especially given inconsistencies in the associated metadata. The new tools speed up the process of turning raw data into models and visualisations by integrating the datasets either

automatically or semi-automatically – although humans still provide input.
The tools have been tested with numerous datasets such as Twitter posts, news, pdf documents, and medical images. A tool to extract loosely structured data in archival documents was developed alongside the Swiss Federal Archives and was used to identify documents without openly accessible data. Collaborations with the cantonal hospitals of Fribourg and Bern resulted in a prototype of a novel architecture to integrate information on prostate cancer, including medical images.

—

Tighten-it-all: big data Integration for loosely-structured data
Philippe Cudré-Mauroux (University of Fribourg)

Data streams: monitoring in real-time

This project developed efficient parallelised algorithms to monitor in real time whether a fast stream of large data complies with specified rules. The more complex the rules, the harder it is to check them efficiently against enormous volumes of data. Moreover, monitoring algorithms have to be scalable for parallelised execution in computer clusters.
The project developed monitoring algorithms for highly expressive and therefore useful input languages, and tested them in two concrete applications. A tool was created to study the auditing of expenses claims and payments. A collaboration with telecom company Huawei produced a proof-of-concept tool to verify that processing is data neutral, i.e. it cannot be controlled by the provider. The new algorithms were shown to compete with industrial-grade monitoring systems.

—

Big data monitoring
David Basin, Dmytro Traytel (ETH Zurich)

Scala programming language: enabling big data analytics

This project introduced several new concepts for Scala, a programming language developed in Switzerland that has become a leader for data science platforms and tools. The team integrated several new technologies into a coherent set of abstractions for interfacing with a large data set, and validated them in open-source projects. The new version of Scala, published in 2021, is expected to be adopted by hundreds of thousands of software developers worldwide.

—

Programming language abstractions for big data
Martin Odersky (EPFL)

Big data analytics

Stream analytics: fast processing and privacy-preserving tools

This project developed tools that contribute to real-time continuous processing of large data streams and privacy protection by adding noise to datasets. The first of these tools is a new algorithm for image processing that was successfully tested with data from the Australian Square Kilometre Array Pathfinder – a radio telescope that generates up to 2 gigabytes of raw data per second. The algorithm could reconstruct astronomical images in real time without having to wait for all of the image data, paving the way for on-the-fly peta-scale analytics.
The second tool helps non-specialists with so-called differential privacy, which randomly modifies data points to avoid the risk of re-identification. The team developed methods to add differential privacy in real time to a continuous data stream and proposed a new way to represent the parameter controlling the amount of added random noise. This allowed users to decide for

themselves the trade-off between privacy (large noise) and accuracy (smaller noise). The team also published a modified programming language to simplify the integration and control of differential privacy techniques for non-specialist users. They tested these concepts on the analysis of TV viewing habits of around 3 million individuals.

—

Privacy preserving, peta-scale stream analytics for domain-experts
Michael Böhlen (University of Zurich)

Coresets: big data with less data

This project showed it is possible to reduce the amount of data needed to train machine learning models while only slightly reducing the accuracy of key statistical analyses and learning processes – something that can also work in dynamical data streams. The team used so-called coresets, which use a small sample to summarise large datasets but which can be processed robustly and accurately. This approach can also improve privacy because individual pieces of data are significantly changed in the coreset samples.

—

Scaling up by scaling down: big ML via small coresets
Andreas Krause (ETH Zurich)

Data centres: efficient performance monitoring

This project devised novel ways to analyse performance in cloud data centres, an important task in managing computing resources efficiently while minimising energy consumption. The team used approximate analytics based on subsets of performance data – logs of virtual and physical computing resources – to predict complex patterns and series of resource usage. This approach could improve current monitoring systems, which tend to be unsophisticated and slow.

The results show that it is more effective to learn from a small set of clean and informative data than it is from a large amount of low-quality data. Choosing the proper subset is essential to avoid losing too much accuracy in the predictions and inferences. This requires a systematic summarising of the data set, instead of taking random or uniform sub-samples. The team developed a method to synthesise tabular data such that proprietary data provided by commercial companies can be shared without being leaked.

—

Dapprox: dependency-ware approximate analytics and processing platforms
Lydia Yiyu Chen (Delft University of Technology, formerly IBM Research Zurich)

Machine learning models: robustness and generalisability

This project made several theoretical advances that help to evaluate if machine learning models are robust – namely, still reliable when the input data is perturbed– and – namely that they can process real-world data distinct from the data used to train the models. The team found a trade-off between a model's efficiency and generalisability. Such advances are important to improve models' interpretability and reproducibility, crucial ingredients in ensuring that algorithms are unbiased and reliable.

—

Theory and methods for accurate and scalable learning machines
Volkan Cevher (EPFL)

Fast prediction algorithms

This project created algorithms that allow learning from large datasets several times faster than state-of-the-art methods. It involves a powerful statistical method called Gaussian processes, which is used to model data, draw inferences and make predictions. Predictions are typically expressed not as single values but as probability distributions, which is particularly important when uncertainties need to be quantified – as is the case in weather forecasts. The team used a Kalman filter formulation over a reduced set of training data, so-called inducing points, and local approximations using a so-called correlated product of experts over the dataset. With the new method, the computing time and complexity scales only proportionally to the size of the data samples, while previous methods scaled cubically. This offers good prospects for scaling to very large datasets. It also allows better estimates of the uncertainty on predictions and inferences.

—

State space Gaussian processes for big data analytics
Marco Zaffalon (Istituto Dalle Molle di studi sull'Intelligenza Artificiale USI-SUPSI)

Language models: new methods for conversational agents

This project made several theoretical advances in the field of language models – systems that generate text given certain inputs – in particular for conversational agents or dialogue systems used to answer queries. This task requires answers adapted to the questions, pointing to a certain understanding of inputs expressed in natural languages such as English or Mandarin. Such systems are widely used in customer support, search engines, social media and e-commerce, with important economic and social consequences.

The team developed technologies that can be integrated directly into dialogue systems. The new methods can be used to understand text elements referring to an agent (entity detection and linking), generate text by deep neural networks, evaluate and improve the performance of language algorithms (reinforcement learning) and of geometric machine learning (which allows complexity and structure to be added to data representation). Part of the work was carried out with Google Zurich. A spin-off company was set up to commercialise solutions for the semantic search of legal files and the redaction of documents for legal teams.

—

Conversational agent for interactive access to information
Thomas Hofmann (ETH Zurich)

4. Societal, legal and ethical aspects of big data

The collection, analysis, storage and sharing of large data sets raise many profound societal, legal and ethical questions. Finding the answers requires interdisciplinary efforts that bring together multiple stakeholders. This chapter reviews the new perspectives brought by NRP 75 on issues of data ownership, control, access and transfer, of privacy and digital sovereignty, of discrimination and fairness, and of knowledge management.

The ever-greater supply and use of data in society has attracted the attention of governments, companies, organisations and individuals, raising profound ethical, legal, and social questions. These issues are widely discussed but have not yet been entirely mapped. Adequate guidelines for collecting, analysing, using and sharing data have yet to be conceptualised, developed, tested, accounted for in governmental and institutional regulations, and put into practice.

The influence of big data can be observed in all aspects of society, such as in social media where the combination of numerous information streams would be unthinkable without advanced analytics. These technologies also enable wilful misconduct and manipulation, generating new risks or exacerbating existing ones at unprecedented scale and speed. Democratic processes are affected, positively and negatively.

The political sensitivity around the potential negative impacts of big data has increased substantially in the last few years. This was visible during the revisions of the New Federal Act on Data Protection, which will come into force on 1 September 2023.

The insights produced by NRP 75

The National Research Programme "Big Data" (NRP 75) has analysed numerous ethical, regulatory and legal issues raised by the rapid growth of big data applications and practices, both at a broad conceptual level and in specific contexts.

NRP 75 researchers studied specific application domains such as healthcare, traced how regulations spread internationally, wrote guidelines for the insurance industry, studied the potential for discrimination in human resources, developed frameworks for the ethical use of data in healthcare and examined how the new profession of data scientist has emerged. They also analysed generic questions related to sovereignty and control of data, regulation of the use of big data in research, and the need to address new uncertainties brought about by predictive models.

The results of the NRP 75 research on the ethical and legal aspects of big data are organised according to four important themes:

- data ownership, control, access and transfer in section 4.1;
- privacy and digital sovereignty in section 4.2;
- fairness, non-discrimination and inclusiveness in section 4.3;
- knowledge production and management in section 4.4.

An outlook on societal issues is presented in section 4.5. The last section 4.6 summarises the eight related NRP 75 projects.

4.1 Data ownership, control, access and transfer

Data is produced, collected, analysed and shared at an unprecedented scale by organisations and individuals with a variety of roles. For example, governments publish part of their data – usually aggregated and rarely at the level of individuals – in keeping with the paradigm of open data. In contrast, commercial entities tend to hoard data, sharing it only reluctantly.

For individuals, personal data has a very special value – and is created on a massive scale. People share and give access to much of their data in exchange for free services that are convenient, efficient and often hard to live without, such as email, messaging, picture and video sharing, maps, targeted recommendations and social media. The public is to some extent aware of the risks to privacy, as well as possible surveillance and misuse, but the very liberal sharing of personal data continues unabated. National and international regulations such as Switzerland's Federal Act on Data Protection or the EU's GDPR (General Data Protection Regulation) are being updated or enacted, but only slowly. Their effectiveness is often limited as they come into force years or decades after data starts being collected.

NRP 75 projects studied issues of data ownership and control, as well as the flow of data across borders (see section 4.6 for more details). They analysed

- several profound legal challenges raised by big data, in particular ownership, protection against undue surveillance and self-incrimination by one's data, as well as the enforcement of legal rights when infractions related to data use occur *(Legal challenges of big data);*
- the risk of harming people whose data is used in research on big data, including issues such as discrimination, interference with privacy and the potential misuse of data *(Regulating big data research);*
- the rules of the World Trade Organization (WTO) and the international trade agreements concluded since 2000 as well as their impact on national big data regulations *(Trade agreements).*

Data ownership and control

Several concepts regarding the attribution of rights to data holders have been discussed in the past. Regarding the concept of data ownership, the normative category of property rights runs into problems because legal provisions usually only encompass physical objects, and not data. Data is not material, it can be copied, it is non-rivalrous (which means it can be used by numerous people at the same time without impairing accessibility or utility) and is also generally non-exclusive (being available to many) – the latter two characteristics being also true of public goods. The idea of data ownership therefore lacks relevance *(Legal challenges of big data, Regulating big data research),* and regulators generally no longer build on the concept of data property.

Assigning intellectual property rights to data is equally difficult. Such rights must provide an exclusive legal position that can be exercised against everybody, but an intangible good must meet the specific requirements of the applicable laws. These conditions are not met for data, as shown by experience in most practical cases. While software is subject to copyright protection, data does not reach the level of a specific creation of an individual mind that is required for assigning intellectual property.

Other legal categories such as neighbouring rights (a special kind of copyright), torts rules (compensations) or personality rights (controlling the commercial use of one's identity) can apply in specific situations. These normative provisions, however, do not supply a suitable basis for a general framework of data ownership.

It is the access to data and its control – not its potential ownership – that determine how it will be used. The people or institutions holding and processing the data are in effect in an ownershipposition and have the power to decide how it is used, stored, deleted and transferred. Consequently, regulators usually build on the concept of data access.

Access to one's data

The design of rules for data access is decisive to those directly or indirectly involved in processing data. Some data controllers voluntarily grant individuals access to their personal data, a position known as philanthropy, but several other types of access rights exist.

— Some general legal instruments specifically address access to one's own data. For instance, the data portability right – part of the New Federal Act on Data Protection, due to come into force in September 2023 – will allow users to transfer the data gathered from one service provider to another.

— Current regulations implementing rights of data access in Switzerland are found mostly in antitrust law, but also partly in unfair competition law. Their application faces many challenges, for example the design of market delineation, whether data is correct and appropriate, and the definition of market power. Antitrust proceedings are also usually expensive and lengthy, while the competition authority often only delivers its decision after the concrete situation has already evolved.

— Data access can be restricted by sector-specific regulations, for example in healthcare.

Internationality

Personal data has become increasingly valuable as economic goods. Transferring data across jurisdictions calls for agreements across borders and harmonised regulations. But this assessment collides with the multitude of national laws framing the collection and use of data, which are influenced by international trade agreements – either bilateral or global *(Trade agreements).*

4.2 Privacy and digital sovereignty

Data privacy and security are among the most debated policy issues raised by big data. Privacy laws govern the collection and processing of personal data, but it is often difficult to distinguish this type of data from the non-personal variety. In particular, different items of data that are not in themselves personal can reveal personal information when combined thanks to the cross-referencing of multiple databases – a process known as deanonymisation. What's more, analytics can potentially generate new insights by exploiting the complex correlations that exist in very large datasets. Taken together, the technological advances in big data raise doubts about the adequacy of traditional principles of data protection.

Several NRP75 projects studied issues of privacy and digital sovereignty in concrete settings. They

— investigated the ethical, legal and

Societal and ethical issues related to big data

Privacy Individuals should be protected against undue access to their private data, and against others sharing and analysing it.
Access People should be able to access and delete their personal data stored by service providers.
User agency Users should be able to control which personal data is collected, how and for which purposes – beyond simply giving or withholding permission for the use of certain cookies.
Societal agency The development of big data is led mainly by corporations, with little control by citizens or public authorities.
Power asymmetry Citizens, companies and governments are, in practice, often unable to change providers.
Regulation Even algorithms bearing great responsibility are largely unregulated, in contrast to physical medical products or vehicles. National regulations differ, hindering transparency and transnational research projects.
Bias Data is not neutral: it reflects existing biases in society, such as limited representation of minorities or correlations of a discriminatory nature.
Fairness Algorithms trained on biased data will likely produce unfair results.
Blackbox The result produced by a machine learning algorithm often cannot be explained. This impairs reliability and trust.
Trust Society needs to have confidence in big data applications. This requires trusting the whole process of data generation and use, in terms of issues surrounding the data itself (privacy, access, and bias) the algorithms (reliability and fairness) and the implementation of big data systems (intentions, ethics, control, etc.).
Innovation Innovation requires clear, stable and balanced regulation.
Business practices Big data applications require data sharing, which raises issues of business confidentiality.

Insights about the ethical, legal and social Issues of big data

NRP 75 put in place the *ELSI Task Force* to study specific ethical, legal and social issues of big data. The key messages of their studies are summarised below.

Big data and digital sovereignty

The concept of digital sovereignty can have two meanings: a state's autonomy in regulating and protecting its citizens' data, or users' self-determination as to how their personal data will be employed. These two meanings conflict with each other because big data is a good over which both individuals and states seek to exercise control. As big data is intangible, governments cannot regulate it based on sovereignty over a finite physical space. Instead they must cooperate with one another to regulate the collection, storage, sharing and transfer of big data. The state's protection of its digital infrastructure must be weighed against its citizens' autonomy, in order to avoid unjustified intrusions into people's private lives.

The challenges of informed consent

It might be more transparent to abandon, at least partly, the need to seek people's informed consent when using their data. If so, such consent should be replaced by mechanisms that aim to ensure data protection through anonymisation and guarantee appropriate compensation if those mechanisms fail. People would receive clear information about how their data is used, while abuses would either be unlikely or detectable early enough to prevent significant harm – regardless of how people want their data to be used.

The importance of guidelines

Data protection regulations are important in protecting individuals' rights, but the heterogeneity of today's legal landscape makes such regulation hard to understand. It is therefore important to find a bridge between the law and practical issues. Given the rapid changes to big data, it may be better for governments to provide guidelines rather than formulate laws, particularly in the case of informed consent.

The ethics of big data in healthcare and biomedical research

Big data could potentially improve healthcare services but raises questions about existing oversight mechanisms such as the use of Ethics Review Committees – in particular whether such bodies should carry out reviews when the data in question is public or anonymised. Clarifying the role of these committees, which should have expertise in areas such as data science, could make researchers more responsible. Complementary bodies such as data ethics boards could also be considered.

The big data divide between companies and customers

There is a divide between those who collect, store, and analyse large datasets, such as companies, and those whose data is collected, the customers. Rather than trying to overcome the divide, it might be better to safeguard those threatened by it – given the divide's near inevitability in our current society. As such, values like non-maleficence and fairness would be more important than autonomy and privacy.

Risks of discrimination against certain groups

Machine learning applications can introduce bias, discrimination and unfairness. Insights needed to combat these problems include the realisation that bias has a normative as well as a technical element and that policy makers should include fairness as a requirement for the design of high-stakes algorithms. Discrimination, meanwhile, cannot be avoided by forbidding processing of group information since such information is needed to assess the fairness of algorithms. Different statistical standards can be used to assess whether an algorithm is biased, and anti-discrimination policies should allow using different standards depending on the concrete cases.

ELSI Task Force of NRP 75, Markus Christen (University of Zurich)
Ethical, legal and social issues of big data – a comprehensive overview, Eleonora Viganò (Ed.), NRP 75, (2022)

societal questions raised by the use of big data in private insurance, and formulated recommendations *(Big data in insurance),*
- mapped the ethical issues of using big data in healthcare, assessed the existing oversight mechanisms and developed both an ethical framework and policy recommendations *(Big data in health),*
- analysed how human resource departments have exploited big data and evaluated the impact of these initiatives on trust among employees *(Big data in human resources).*

Principles of data protection

One fundamental issue is the question of which people or institutions decide how data is used in different settings. The concept of digital sovereignty is often used in this context, although the term is ambiguous: it can mean informational self-determination but can also mean personal control over one's data as well as the idea that data collected within one country should be subject only to the laws of that state and not to those of other states. A thorough analysis of such sovereignty is therefore important for future regulations (see the NRP 75 white paper *Ethical, legal and social Issues of big data – a comprehensive overview*).

Most data protection laws involve the principle of data minimisation, namely limiting the collecting, processing and use of data to what is required for a specific purpose. In principle, this idea aims to minimise the indiscriminate accumulation of information for additional uses, whether foreseen or unforeseen. However, huge datasets are usually designed to find correlations and generate new information. The fact such results are unforeseen can be used to justify the collection and processing of data, since consent by its very nature can only be given regarding a procedure with known outcomes. A parliamentary motion for drafting a law on the secondary use of data was submitted on 22 August 2022[10].

The principle of informed consent – that users must give their explicit and informed approval to the collection, processing, use and transfer of their data – is not easy to apply in practice. Indeed, most individuals concerned are not in a position to assess the contractual context of data processing or to evaluate the risks associated with big data analytics. Since the traditional data protection principles are ill-suited to big data, there is an urgent need to rethink these principles and laws.

New governance models and privacy-enhancing technologies

Forward-looking privacy concepts should include governance elements such as:
- adequate analysis of the relevant and potential risks to data protection,
- establishment of an appropriate strategy to comply with principles of data protection,
- application of data protection policies,
- introduction of appropriate procedures to remedy failures of data protection.

Privacy can be protected not only through regulation but also via technology such as end-to-end cryptography or techniques that ensure anonymity. Suitable infrastructures, such as the ETH Zurich Polybox ecosystem,

[10] Rahmengesetz für die Sekundärnutzung von Daten, Motion 22.3890 (2022)

can guarantee that data is shared securely while also ensuring that data are findable, accessible, interoperable, and reusable *(Big data in health).*

4.3 Fairness, non-discrimination and inclusiveness

The issues of fairness, non-discrimination and inclusiveness have gained much attention as regards modern technologies. As with privacy, the risks are particularly acute for socially sensitive segments. The ethical implications of big data must be understood both broadly and within the context of specific applications, such as those in insurance, the workplace or healthcare.

Data fairness in concrete applications

Insurance industry

Increasingly detailed data about insured individuals enables more personalised underwriting. This ultimately threatens the ability of insurers to pool risks, so undermining the principle of solidarity that underpins the concept of insurance *(Big data in insurance).* Indeed, it raises the provocative question: does big data render people uninsurable? It is not impossible that some people may in effect be excluded from the insurance market because they are considered too risky. Ethical principles play an important role here, as well as legal regulations. Specific codes of conduct could be developed for insurers when using big data analytics (see “Insights for the insurance industry”, p. 59).

Employment relations

Employers must adhere to the principles of non-discrimination and fairness and mitigate the effects of discriminatory algorithms. Human integrity is particularly important in employers’ relationship with their employees, and for building and maintaining trust in the workplace *(Big data in human resources).*

Staff monitoring should not compromise integrity or moral reputation. Employers must communicate with their employees transparently and involve them in the design of their environment. Human resources is a particularly sensitive area, since its activities can potentially violate employees’ rights or harm their reputation.

Healthcare

The concepts of informed consent, minimal risk and privacy are particularly important as regards health *(Big data in health, Regulating big data research).* Beyond the prescription of data compliance, guidelines and toolkits should ensure uses of data that are ethical.

Research on big data for biomedicine and beyond involves a variety of stakeholders, including researchers, developers, the private sector, professional organisations, decision makers and the public, each of whom might have competing interests and different expectations about sound healthcare. As such, existing oversight mechanisms should be reformed and strengthened at different levels, and adapted to the new technology.

4.4 Knowledge production and management

The creators of big data applications play a central role in developing and maintaining ethical guidelines and practices, both general and domain specific. In addition to the project *Regulating big data research*, two other NRP 75 projects studied how big data impacts the research and knowledge professions.

- One team followed epistemological approaches to evaluate how big data is used in computer simulations for scientific research, particularly those in climate science *(Uncertainty in big data).*
- One project used methods from sociology and ethnology to study the role of big data in sociology, data science and data journalism *(Big data in practice).*

Transdisciplinarity and new skills

The legal, ethical, and social questions surrounding big data also extend to the natural sciences. As such, interdisciplinary approaches are needed to contend with large and novel datasets.

The inclusion of persons in matters that affect them must be improved. Inclusiveness is part of a fundamental change in the production of scientific knowledge as well as in the understanding of the structures and mechanisms involved in maintaining and changing knowledge domains. Data and visual literacy as well as computational thinking competences are crucial for extended participation in knowledge production endeavours *(Big data in practice).*

Context matters

The variety of research on big data means ethical issues cannot be addressed by overarching one-size-fits-all regulation. Context and deliberation should be emphasised, not inflexible standardisation *(Regulating big data research).* The interdependency of legal, ethical, and social issues, and the involvement of different cultures, calls for interdisciplinary, international research.

Understanding big data knowledge

Science employs big data analytics in a variety of ways, including synthesising data points, making predictions and discovering relationships. The results of these analyses come with numerous uncertainties, which must be assessed, quantified and properly communicated if the results are to be trusted and used *(Uncertainty in big data).*

4.5 Challenges and key messages

Challenges in investigating the societal, ethical and legal issues of big data

Research on big data in society has to deal with rapidly changing technology and regulatory issues, such as the dwindling concept of data ownership *(Legal challenges of big data).* This calls for flexible research programmes and funding.

As with projects on applications and infrastructure, it can be hard acquiring data (even academic data) for this social-science research *(Big data in practice, Big data in insurance Industry,* *Uncertainty in big data)* – although it was possible in one area *(Trade agreements).*

Ethical and legal evaluation of big data systems is essentially interdisciplinary, which makes it hard to agree on terminology, methodology and concepts – particularly among data scientists, modellers and legal scholars *(Legal challenges of big data, Uncertainty in big data).*

Key messages

Although NRP 75 only partially addressed a subset of all societal, legal and ethical questions relating to big data, its research nevertheless highlights some generic issues.

- Both the public and private spheres must be more transparent about the use of big data.
- Additional academic research should consider big data's potential impact on democracy. These include the increasingly important role of analytics in social media, which accelerates the dissemination of false information and data manipulation. Such developments should not be left to the discretion of commercial companies.
- Aside from the technology, societal aspects of big data analytics should not be neglected in future deliberations.

Addressing societal issues

New data technologies are disrupting personal and business life, with data aggregation and analysis as well as analytics techniques having major effects on sectors such as healthcare and the workplace. This requires carrying out contextualised analyses and anticipating social consequences, as well as drafting practical guidelines adapted to different environments *(Big data in health, Big data in human resources, Big data in insurance, Regulating big data research).*

Drafting appropriate regulation

The legal system needs to set up a proportionate normative framework. Given that the concept of ownership is ill-suited to (non-physical) data, legislators could start to formulate the alternative concept of a data right holder that centres on control of and access to data. New normative instruments will be needed in areas such as the blockchain, and there should be neither too much nor too little regulation.

Developing ethical guidelines

Ethical guidelines are needed because inclusiveness and fairness are not well

Insights for the insurance industry

The NRP 75 project *Big data in insurance* analysed the ethical and legal issues raised by big data analytics in personal insurance. It formulated several concrete suggestions for the industry and the regulating authorities:

→ Whether and how insurance companies should be allowed to personalise their insurance contracts based on big data analytics are questions that should not be resolved indirectly by applying the general data protection and anti-discrimination laws.

→ The Swiss regulator should continuously monitor and anticipate the use of big data to personalise insurance contracts. Insurance law should be amended to prohibit unwanted forms of personalisation or define what type of personalisation is permitted.

→ Insurance companies should avoid using data sources unrelated to the insured risk, so as not to undermine customers' trust in the industry's products and services.

→ They should be aware of how discriminatory use of machine learning can affect prediction, pricing and fraud detection.

→ They should show their clients how they protect privacy, fairness and solidarity when using big data analytics.

→ They should adapt their ethical principles so that they are accountable when dealing with issues raised by the industry's digitalisation.

Big data ethics recommendations for the insurance industry, NRP 75 (2019) prescribed in law while many principles of non-discrimination are not covered by constitutions.

Such guidelines should therefore be created using concrete processes that are situation-specific and involve multiple stakeholders – the more stakeholders there are, the more likely it is that the guidelines will be followed. *(Big data in health, Big data in insurance, Regulating big data research).*

4.6 The research projects on societal, legal and ethical aspects of big data

Societal issues

Big data in practice: sociology, data sciences and journalism

This project used sociology and ethnography to analyse how big data is understood, taught, or used in sociology, data sciences and data journalism. It analysed more than 750 syllabi from German universities and identified four cultures in teaching sociology methods.

To understand how data sciences are perceived and directed in Switzerland, the team analysed some 4 300 online job advertisements, 34 policy and strategy papers as well as 40 new curricula in Swiss higher education institutions. The results show that data sciences share a distinct set of methods, tools and practices, that they transcend boundaries between disciplines while lying on the front line of conflicts between say computer science and statistics.

Ethnographic studies show that data journalism requires specific epistemological and professional cultures to be coordinated with existing journalistic practices. Overall, the results underline that big data calls for skill sets that cross disciplinary boundaries.

—

Facing big data: methods and skills needed for a 21st century sociology
Sophie Mützel (University of Lucerne)

Uncertainty in big data applications: lessons from climate simulations

This project used epistemology to study computer simulations developed for climate research and based on big data. It shows that climate researchers are using ever more diverse data, such as social media or web searches, which makes computation more efficient and uncovers new relationships within models. However, uncertainties arise from variations in data quality and incomplete understanding of the data's role. With research results compared regularly against new observations, the project underlines the importance of combining big data methods with traditional scientific approaches, such as process understanding. It calls for transdisciplinary collaborations between subject experts and data scientists, particularly to assess uncertainty.

The team analysed two climate case studies: high-resolution modelling and predicting of urban heat islands, and urban temperature dependence on vegetation and other factors. In the first case, the research differentiates between uncertainties based on limitations of the prediction algorithm itself and those derived from finite training data. The latter affects specific objects more than it does generic categories of object.

—

Combining theory with big data? The case of uncertainty in prediction of trends in extreme weather and impacts
Reto Knutti (ETH Zurich)

Legal issues

Legal challenges of big data

The project explored several legal issues raised by big data. It found that data ownership cannot currently be defined, and that markets for personal data such as targeted advertising operate essentially without legal regulation. The researchers evaluated possible alternatives to existing normative concepts regarding data rights.

One legal analysis indicates that agreeing to share personal data with a service provider only partially waives rights of data privacy. Another study underlines the conflict between the authorities' mandate to protect society and the citizen's fundamental right not to be unduly surveilled. For example, car data allows forensic reconstruction of accidents but may violate basic legal principles such as the privilege against self-incrimination. The research discusses numerous unresolved issues regarding data rights, such as clarifying who are the victims of data crimes, and the fact that citizens of Switzerland and Germany currently struggle to claim legitimate interests in criminal proceedings.

—

Legal challenges in big data. Allocating benefits. Averting risks
Sabine Gless (University of Basel)

Trade agreements: impacts on national law

The project looked at hundreds of trade agreements concluded over the last two decades for their relevance to the data-driven economy. It analysed existing norms and the ever more numerous provisions directly relevant to big data, such as those on electronic commerce, data protection and open government data.

The team analysed the interplay between international commitments and national policies, and also created an openly accessible database now employed by the OECD, the UK Department of Trade, the WEF and others. The research found that data regulation requires more international cooperation, with policies in areas such as data protection and national security varying significantly across countries. The results highlight the growing importance of trade law and suggest ways of better using such law in data-driven economies. The project argues that Switzerland could play an important role as an innovative and globally connected country.

—

The governance of big data in trade agreements: design, diffusion and implications
Mira Burri (University of Lucerne)

Regulating big data research

This project analysed the many ethical issues raised by research with big data, in particular the problems of discrimination, privacy violation and data misuse. While ethical and regulatory processes are well known in research on humans, such as medicine or psychology, the use of anonymous data in science also raises many (little-known) issues.

The project found that the varied nature of scientific work makes it hard to build a comprehensive, harmonised and standardised framework for research on big data. It suggests that regulation should instead involve decisions based on context, ethical deliberation and analysis of trade-offs – a potentially ongoing process. Review boards should be made up of professionals including big data specialists, who would assess the ethics of research projects throughout their lifecycle. Ethical assessment must also cover research performed by private companies, which are increasingly collaborating with academia. Guidelines, procedures, and codes of conducts will have to be revised regularly to keep pace with changes in technology and regulation, such as the development of algorithms questioning the efficiency of anonymisation or the implementation of Europe’s GDPR regulation.

—

Ethical and legal regulation of big data research – Towards a sensible and efficient use of electronic health records and social media data
Bernice Simone Elger (University of Basel)

Ethical issues

Big data in health: an ethical framework

This project looked at the ethics of big data in healthcare, assessing whether such issues could be dealt with effectively by existing oversight mechanisms like Research Ethics Committees. It also developed an ethical framework and policy recommendations to support these mechanisms, proposing a practical toolkit for researchers and ethics committees.

The research shows that scientists and app developers tend to see big data ethics as compliance with existing data protection regulations, while ignoring issues such as research accountability, fairness, individual autonomy and harm to groups. Ethics committees in Switzerland have admitted they lack expertise in big data and have asked for training. The project concludes that ethics committees must change their regulations and procedures to help them oversee biomedical research. The team developed a toolkit to help ethics committees assess their readiness for big data research and a checklist to facilitate project reviews.

—

BEHALF – Big-data-ethics-health framework
Effy Vayena (ETH Zurich)

Big data in insurance

An interdisciplinary team investigated the ethical, legal and societal issues of big data in private insurance, formulating brief recommendations for insurance companies with large client datasets (see “Insights for the insurance industry”, p. 59).

The project found that privacy is becoming less of an issue than predictive analytics, which quantify risk but also reveal clients’ propensity to pay premiums or engage in fraud. The increasing granularity of risk assessment makes discrimination against specific behaviour (such as not practising any sport) more likely, while also indirectly discriminating against the population groups most given to that behaviour. Analysis of legislation shows that Switzerland is rather liberal and respects the principle of freedom of contract much more than say California.

—

Between solidarity and personalization – Dealing with ethical and legal big data challenges in the insurance industry
Markus Christen (University of Zurich)

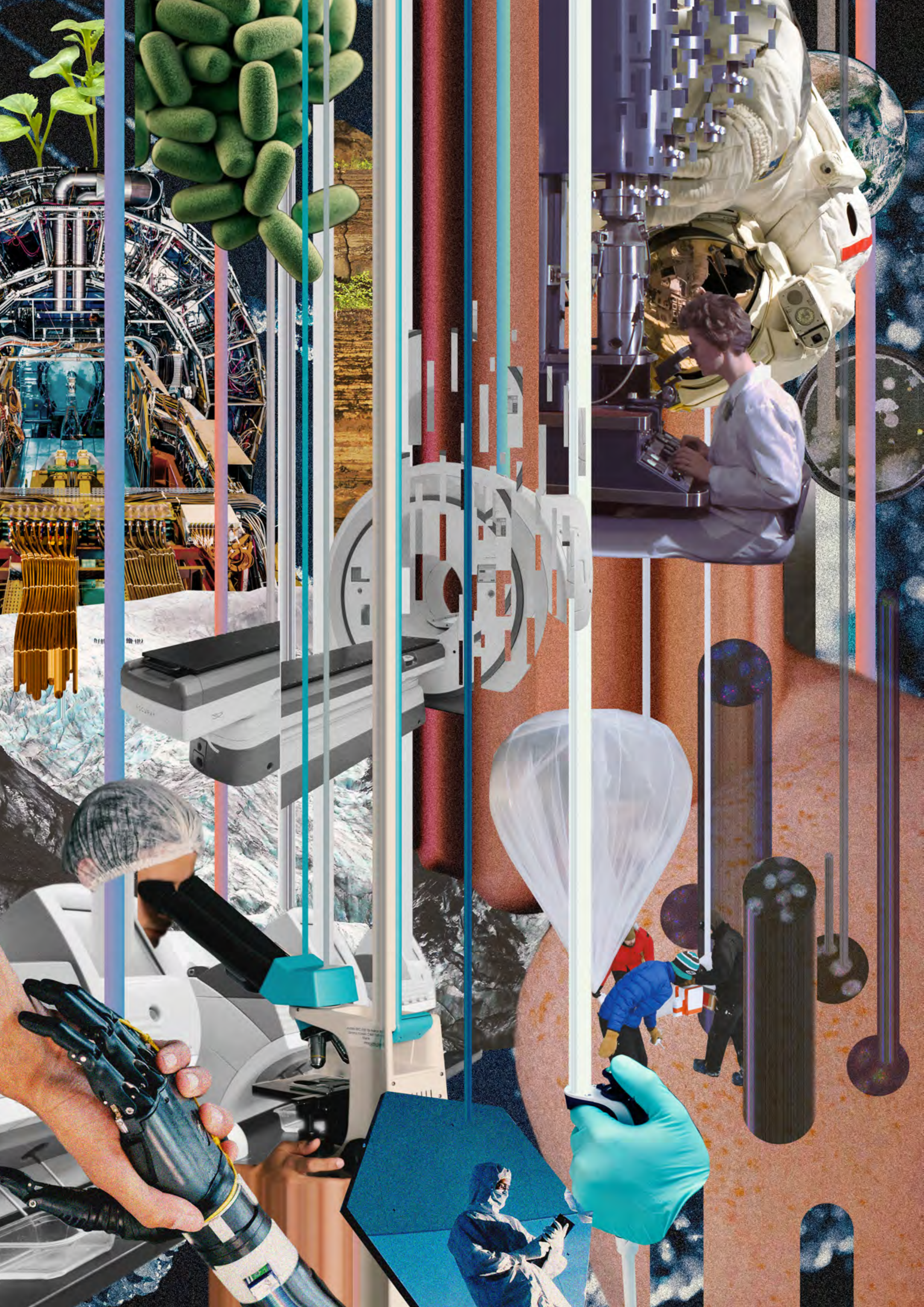
Big data in human resources

This project analysed how human resource departments, particularly in Switzerland, have integrated big data, and how that integration has affected trust between employees and employers. It uncovered much variation in firms’ transparency of data collection and empowerment of employees, for example, which in turn led to different levels of trust in the employers. Building trust requires protecting staff’s autonomy and agency as well as safeguarding privacy, transparency and control. Staff must be involved in strategic decisions and their grievances listened to.

The results show that Swiss laws in particular are ill-equipped to deal with discriminatory algorithmic systems in the workplace, highlighting the need for appropriate legal safeguards and modernisation. The team developed a toolkit to familiarise HR managers with big data technology and enable evaluations of legal, ethical and workplace issues.

—

Big data or big brother? Big data HR control practices and employee trust
Antoinette Weibel (University of St Gallen)



5. The road ahead

Society must anticipate the potentially disruptive changes that big data and machine learning applications can bring about. An overview of key opportunities and challenges that lay ahead is given below.

The main achievements of the National Research Programme "Big Data" (NRP 75) are having reinforced Swiss capabilities in the technologies, applications, and societal aspects of big data. NRP 75 has advanced technologies that underpin big data infrastructure and has brought together big data scientists and domain experts to realise specific applications. It has also raised awareness of the societal challenges that accompany large-scale data production and analysis, and has helped develop a "big data culture" to reap the benefits of big data responsibly.

The programme's 37 funded projects covered only a subset of this rapidly expanding field. This chapter goes beyond to provide a more general view of the opportunities and risks that big data presents, particularly those that could become more prominent in the coming years. The following analysis is based on the knowledge gained from NRP 75's research and on the collective insights of the programme's Steering Committee members. It addresses both the prospects of a greater use of big data in industry and the public sector, and the challenges of sustainability, privacy and accountability.

5.1 Applications impact more domains

Many more applications of big data are expected to be developed and deployed in the coming years. New private sectors – beyond e-commerce – and public administrations will become data-ready in the hope of staying competitive, by developing new capabilities and reducing costs while increasing efficiency. As shown by NRP 75 research projects, developing real-world applications requires the right combination of expertise across several domains. It needs robust data strategies including privacy-by-design approaches, analytic know-how among sector experts, and knowledge among the work forces. A crucial ingredient is the availability of data scientists who understand the relevant application domain and domain specialists familiar with data science. This underlines the importance of equipping the new – and older – generations with the knowledge and tools needed to tackle big data applications (see conclusion 1 in chapter 6).

Here follows a selection of domains that could be strongly affected by big data applications.

Production: improving output, optimising maintenance

Many manufactured products include integrated sensors that can be connected to the Internet of things (IOT). Products can send performance information in real time, allowing manufacturers to identify components that need replacing or improving, or boosting safety and customer satisfaction.

In agriculture, autonomous robotic systems use image recognition to remove weeds, detect diseases and pests, harvest fruit, apply fertiliser locally, and monitor whole fields with drones. Such robots could help reduce labour shortages, lower fertiliser use and avoid pesticide[11].

[11] See e.g. https://www.agricultural-robotics.com for an overview.

State: improving infrastructures and supporting the energy transition

Governments can use big data to enact evidence-based policy making, such as allocating resources, carrying out strategic planning or monitoring public infrastructure (conclusion 5). Analysis of big data can improve transport planning *(Optimising transport management);* manage traffic congestion; improve the planning, construction and operation of utilities providing water, electricity, and lighting, for example; and carry out environmental monitoring *(Soil erosion, Flood detection).* Sophisticated analytics will help reduce our carbon footprint by ensuring flexible energy supply, storage and distribution, and in particular enable electricity grids to handle decentralised and intermittent renewable sources of energy such as solar panels or wind turbines *(Renewable energy potential).*

Services: automating actions in finance and cybersecurity

Financial institutions can use real-time transaction analytics and market predictions for fast automated trading, which needs efficient infrastructures *(Fast prediction algorithms, Graph analytics and mining).* Quantifying individual risks allow insurance companies to better tailor their policies, but potentially threatens the principle of solidarity that underlies insurance *(Big data in insurance).* Tracking systems in vehicles and elsewhere could reward risk-mitigating behaviour, so putting the emphasis on risk prevention rather than risk protection.

Analytics can help prevent cyberattacks by looking for anomalies in real time data transfer and then automatically blocking threats *(Data streams).* Image recognition can be used to automatically detect physical security breaches and other irregularities.

Healthcare: assisting health professionals and personalising medicine

It is widely expected that machine learning will significantly improve healthcare *(conclusion 4),* having already been used to identify anomalies in clinical imaging[12]. New technologies could enable major progress in prevention, diagnoses, and targeted therapies by bringing together huge datasets from laboratory testing, health records and genetics. In particular, advanced natural language processing *(Language models)* allows automatic extraction and interpretation of information from unstructured texts in health records. By integrating data streams from various clinical devices in real time it is also possible to measure patients' health and detect emergencies *(Intensive care units).*

However, using big data for medical applications requires certain infrastructure. In particular, innovative methods are needed to to extract reliable results from small subsets of data – given that a single patient can generate terabytes-worth. For genomic data, this can be done through adequate pre-processing *(Big genetic data).*

E-commerce and entertainment: consumer involvement and synthetic art

Collecting, analysing and exploiting customer information will likely play an

[12] See e.g. https://grand-challenge.org/aiforradiology for an overview.

increasing role in e-commerce. Online companies already use personalised recommendations and trend predictions, but new data-driven applications will probably incorporate customer expectations into the process of product design itself.

Language models are improving very quickly, getting better at comprehending meaning, intention and context, and extracting valuable information from text, as well as generating synthetic reports or conversations by chatbots. Algorithms can generate music based on the styles of specific composers. Computers using text prompts create convincing-looking synthetic images and videos and software are soon expected to be able to generate films indistinguishable from the real thing – complete with natural-looking people and settings. Such systems may complement or replace current media and entertainment products, but also pose major challenges for democracy by enabling realistic computer-generated image, audio and video hoaxes, but also for intellectual property[13].

Open research: accelerating discoveries

Scientists are increasingly making their datasets available to others for free, in order to accelerate discovery and enhance reproducibility (conclusion 6). But like any other data repository, those supporting open research must comply with certain standards – such as the "FAIR principles" of findability, accessibility, interoperability, and reusability. These call for machine-readable, standardised metadata containing the necessary explanations and descriptions – all part of a new paradigm that the academic world must adapt to *(Big data: open data and legal strings).*

5.2 Decreasing the footprint of big data

While big data will certainly play an important role in tackling climate change and reducing our carbon footprint, it also contributes to the problem. Storing and processing large datasets utilises substantial energy: 3.6% of Switzerland's total electricity consumption in 2019 was due to data centres, a rise of 30% in 6 years[14].

Managing big data is more than mere collection and storage; the data must also be protected from unauthorised access, corruption and loss. This requires access control, backup protocols and solutions to correct damaged, incomplete or inaccurate data. Databases must be preserved by being continuously adapted to new standards of storage, compression and analysis. This requires the work of data and domain experts, and adds to the costs of big data applications. Frugal, or lightweight, artificial intelligence aims to reduce energy consumption, for example by being able to work with smaller datasets and using synthetic training data that save on resources. This new and growing field calls for further research effort *(Coresets).*

[13] The lawsuit that could rewrite the rules of AI copyright, The Verge (2022)

[14] This corresponds to 2.1 TWh, or a quarter of the Gösgen nuclear power plant's production. See "Stromverbrauch der Rechenzentren in der Schweiz steigt weiter an", Swiss Federal Office of Energy (2021).

Ensuring representative big data research

NRP 75 led a programme to strengthen the community of female scientists active in big data research in Switzerland, given that only 22 % of graduates in technical subjects are female – one of the lowest rates across all OECD countries[15]. This situation reinforces the lack of specialists, fosters research topics that are not representative of society, and frames the issues in a biased way. For female experts, this causes them to lack inspiration, encouragement and support, as well as making harassment and discrimination more likely.

The NRP75 cross-cutting activity *Women in Big Data* focussed do both technical and social sciences. It launched various actions to promote careers, such as helping build networks, addressing obstacles to technical excellence, and fostering interdisciplinary exchanges about problems of gender in big data.

Women in big data
Lydia Yiyu Chen (Delft University of Technology)

5.3 Balancing privacy

Numerous big data applications, such as those used in finance, engineering or environmental monitoring, do not raise new questions about privacy as they do not use personal information. But many other applications do, and the ever-growing amount of data they collect about individuals raises ethical and legal concerns. Users generally have little idea about what data of theirs is being collected, who can access it and to what end. The fact that online service providers control these things has led to the concept of digital divide and asymmetry.

Although providers are currently obliged to notify users and ask for consent when they collect data, these steps are not enough to protect privacy since most users agree without further thought and with little knowledge of the consequences. The main problem is that users shoulder the burden of understanding the implication of their consent, even though they gain no immediate benefit from the data collection. Authorities will have to decide how much to regulate this practice (conclusion 8).

Complete anonymisation is often unattainable

Until recently, it was considered safe to share data that contained information on individuals once it had been anonymised – by removing information that could directly identify someone, such as their name, birth date and address. However, it is has become increasingly clear that linking data from different sources, even though anonymised, can enable individuals to be re-identified of. Certain types of data such as whole genomes or a smartphone's GPS traces contain such a high level of sensitive personal information that absolute anonymisation is not realistic. The release of data with personal identifiable information removed must therefore be considered a continuum, requiring that the privacy lost be balanced against the value created on a case-by-case basis.

Several approaches can hinder re-identification. Differential privacy, for example, obfuscates data by adding random

[15] Switzerland: 22% of graduates in STEM disciplines (science, technology, engineering and mathematics) are female, compared to 26% in Germany, 32% in France and 40% in Italy. Computer sciences has an even lower rate of only 16%. Geschlechterunterschiede in MINT-Studiengängen: Eine deskriptive Analyse, KOF, ETH Zurich (2020)

noise, but at the expense of accuracy *(Stream analytics).* Another option is to suppress certain data points or combine them into broader categories, as is done in so-called k-anonymity.

Analysing data without accessing it

Sensitive data can be stored in enclaves with sophisticated control of access. This ensures that only local analyses can be performed, and that only aggregated results, which protect privacy, are sent outside the enclaves. Another option being developed is federated analytics, where the data is kept in multiple local systems with no sharing. The computations, including the training of machine-learning algorithms, are performed locally and collaboratively. Here too the only things shared are partial and aggregated results or intermediate model updates, while the original data is never distributed. This helps solve difficult issues of cross-border data transfer, which requires legal solutions at the international level (conclusion 9). Research teams developing applications for big data should consider the ethical and legal framework of data processing early on (conclusion 2).

5.4 Making algorithms accountable

Big data applications often employ machine learning algorithms able to make predictions based on models trained with certain data. While these algorithms are often very good at predicting, it is often unclear exactly how they arrived at the predictions. This problem can raise ethical and legal questions, as addressed in chapter 4 and in the white paper *Ethical, legal and social issues of big data – a comprehensive overview.*

The risk of discrimination

Normal software follows a strict series of instructions that have been (largely) designed by humans. Programmers and testers can, in principle, guarantee that it works as expected. But the situation is different with many machine-learning algorithms: their results are based on models with huge numbers of parameters, whose values are generated automatically from training data. Their behaviour does not follow human-coded rules.

This makes it difficult to work out whether such results comply with established ethical standards or if they might, for example, discriminate against certain population groups. This may happen if the training data is non-representative, biased, outdated or erroneous, which can be the case when using data from the web. Machine learning models are dependent on training data, so their results may reproduce biases within that data. For example, removing the parameter “gender” from the training data might not prevent discriminatory results, since a trained model might use proxies from other inputs to recreate the gender category. Such behaviour can escape detection during early testing but emerge later on.

Understanding machine learning

As already stated, results produced by deep neural networks and other machine learning techniques can be very difficult for humans to understand because the billions of trainable parameters that make up their models obfuscate the mechanisms leading to particular results. There is currently no accepted solution to fully overcome this “black box” problem of artificial intelligence.

Theorists are trying to better understand these automated systems in order to improve the explainability and traceability of their decisions. These are crucial aims in demonstrating that algorithms are non-discriminatory, accountable and trustworthy.

A typical person or company affected by a potentially biased algorithm has neither the knowledge nor ability to convincingly argue that the system has made a mistake or discriminated against them. One possibility is to reverse the burden of proof, so that the entity responsible for the system has to demonstrate that it behaves correctly. This could involve a certification process developed by a public or private organisation (conclusion 3), and in particular the deliberate alteration of test datasets to see whether the output complies with ethical regulations.

Who is responsible for the algorithms?

The rapid advance of machine learning raises the question of liability, as widely discussed for self-driving vehicles. Who, in that case, should be held responsible for an accident? The owner of the vehicle? The manufacturer? Nobody? This is an evolving area of law and policy, and there is currently no agreement on the answers to these types of questions. While manufacturers must design their cars to minimise risks in typical driving situations, they cannot foresee all possible circumstances. It is essential to define responsibilities precisely so that legal uncertainty doesn’t obstruct innovation.

Towards new regulation

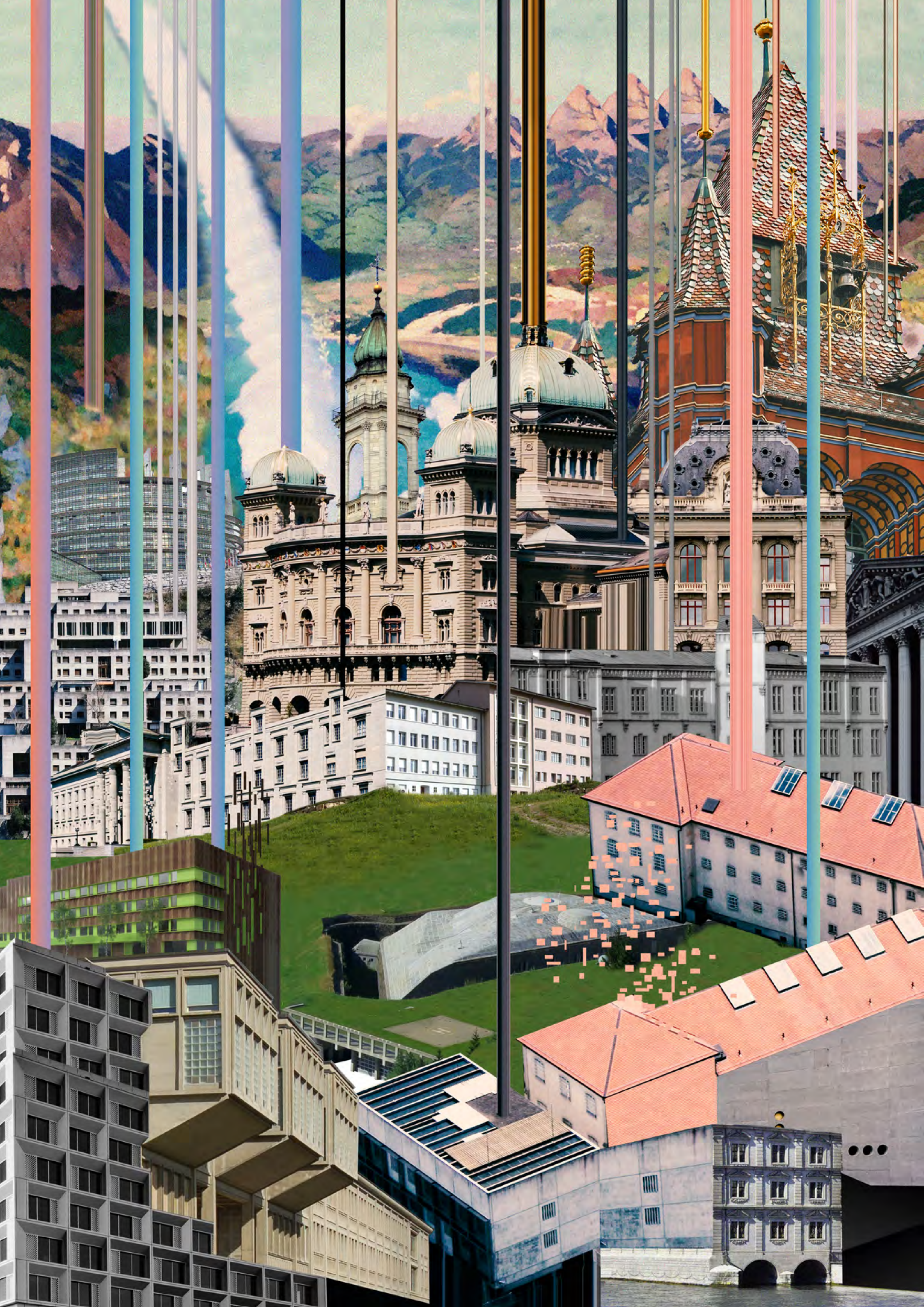
The law often lags behind rapid advances in machine learning and ever-expanding data collection. Legislation has so far focused on individual rights and avoiding negative effects on individuals, rather than impacts on society as a whole.

The EU is currently drafting an act[16] to regulate AI applications. It would ban applications considered unacceptably risky, such as manipulative algorithms or social scoring systems, while restrict those thought to be high risk, such as those managing critical infrastructure or law enforcement. China has also formulated an ethics policy for AI, favouring social security over individual rights[17]. This policy excludes the public sector, which is free to carry out facial recognition and social profiling.

The rapid evolution of technology, driven largely by international companies, poses a difficult problem for the law. Switzerland should proactively draft legislation (conclusion 7), ensuring that rules can be applied concretely and that compliance is monitored.

[16] Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, European Commission (2021); see also https://artificialintelligenceact.eu/.

[17] New generation artificial intelligence code of ethics, Ministry of science and technology (2021)



6. Conclusions of the Steering Committee

The Steering Committee of NRP 75 has formulated nine conclusions on the subject matter of the research programme. Their aim is to provide useful input to public institutions, private organisations and society at large in forming opinions on the developments needed to leverage big data and ensure its responsible use.

Together, big data, artificial intelligence and machine learning will have a profound impact, providing many potential benefits across all sectors of society. The National Research Programme “Big Data” explored a variety of ways to speed up development of new technologies and applications, and address the related societal issues. Staying in control of this evolution represents a major challenge to our public and private institutions. It requires focused efforts in many directions, including education, regulation, sector-wide initiatives and public discussions. In the following, the conclusions of the NRP 75 Steering Committee are presented, based on the knowledge gained from five years of research and on their collective insights.

The conclusions are intended to promote appropriate developments and to support measures already underway, but it is up to stakeholders to decide whether to transform them into concrete actions. They articulate needs and prospects identified from the perspective of research in data analytics and other big data related research fields. This condensate has been elaborated by the members of the Steering Committee based on the outcomes of the research projects and on their own expert knowledge and experience.

Research can produce answers to individual questions and develop specific solutions. However, this might also lead to overlaps or incompatibilities between the different approaches. It is not up to researchers to determine the societal priorities and balance; it is a matter for politicians and voters.

This Programme Résumé with its conclusions constitute a contribution by the scientific community to the formation of opinions, to political and professional debates, and to the planning of strategies and measures to develop big data applications and regulations. It is in particular addressed to actors who play a major role in defining the Swiss data space and who are therefore in a position to shape it.

Foster an appropriate environment for big data development

1
Enhance education of big data professionals

The competent use of big data technologies calls for new knowledge and skills. Today's IT professionals, even those educated a decade ago, may struggle with some aspects of big data such as responsible use of data, data integration and engineering, analytics, machine learning, and visualisation. There is a shortage of qualified professionals all along the value chain and in the respective fields. There is fierce competition for the best talent among academia and large IT companies and start ups. To achieve the benefits of big data in businesses and industry, in society, and in research, it is recommended that school- and university-level education in big data be expanded, such as in the form of enhanced programmes and continued education.

The limited availability of skilled IT professionals is a bottleneck to the exploitation of big data technologies and applications. This calls for better and broader training and education. Only through education, research and outreach activities can Switzerland draw national and international talent to the field, further strengthening its already strong position in big data.

Universities, including universities of applied sciences, should enrol more IT students and offer tailored programmes in big data at the bachelor, master, and doctoral levels. Courses on pertinent aspects of big data should be made available in disciplines that make frequent use of big data. This may increase gender diversity in IT since big data is broader than core computer science, requiring interdisciplinary skills that also encompass societal, commercial, and legal aspects.

Continued educational offerings on big data should be expanded for IT professionals and apprentices, covering all aspects of the big data pipeline. Expertise in data collection and preparation is needed, including skills in integrating, cleaning, and linking data. Further, there is a need for skilled use of infrastructures for storing, managing, and extracting value from data, which calls for training on the rapidly evolving open-source libraries available for these purposes. It is important to master tools that visualise the outcomes of data analyses. And finally, professionals must understand the legal, regulatory, and ethical issues surrounding big data. This can be achieved by offering interdisciplinary programmes.

Since everyone in society is affected by the rapid developments in big data, education in big data must also be intensified and continuously modernised in schools, high schools, and professional apprenticeships. It would cover data science and its various facets, including law, ethics, and societal issues.

2
Support legal and ethical advice for big data research and development projects

Many research and development projects need legal and ethical advice on aspects such as which data can be used and shared, and through which processes. Also, in case a public debate occurs in relation to a project and its findings, the project managers may be challenged to successfully communicate and convincingly demonstrate that its efforts are lawful and ethically sound. This can and should be enabled by making competent consulting and trusted audits available at affordable cost.

Investigators in research and development projects will increasingly need legal or ethical insight to decide which data can be used or shared and to design appropriate services. If neither the project participants, their immediate peers nor their organisation have the necessary legal and ethical expertise, the cautious decision is to not use or share available data – even if the use or sharing would be valuable and perfectly legal and ethical. As a result, a gap opens between real-world settings and the investigated experimental settings, with negative consequences for example in areas such as personalised medicine, public health, or sustainability.

This type of problem is particularly acute at universities, public research institutions, and applied science institutes, as research opportunities are lost, and education programmes become less attractive. The situation is further exacerbated by the risk of negative media coverage and limited protection of university employees. Project managers need a service, either in-house or within the public administration, that provides them with competent, trustworthy, dependable, and quotable advice on legal and ethical issues as well as on communication.

How should activities be designed to be legally compliant and ethically sound? What are the reasons for possible limitations? And how can activities be communicated transparently, particularly in case of contentious public debate? A service could provide advice on such questions, easy-to-follow guidelines for practical implementation, and easy-to-understand explanations of legal and ethical considerations. In addition, it should offer audits of data usage and sharing that address the specific challenges of a project, carried out by qualified, possibly specifically certified, auditors.

3
Enable certification of big data application properties

Big data applications have the potential to improve a wide range of processes, both in public administration and in the private sector. In some cases, applications face concerns related to fairness and biases, discrimination, ethical standards, privacy, etc. To enable the adoption of such applications, it is recommended to provide means of certifying relevant properties of such applications. This includes both the specifying pertinent properties and offering procedures that allow certification of compliance with these properties.

The functioning and outcomes of big data applications often lack transparency, which can generate mistrust. This can be countered on the one hand by defining pertinent properties of big data applications such as fairness, bias, discrimination, explainability, transparency, and accountability. And on the other hand by establishing means that enable application providers to show that their applications possess such properties. To build reasonable trust in big data applications, and thus to enable more widespread use and increase their benefits, it is recommended to establish means to certify relevant application properties.

Numerous applications of big data are already found in non-sensitive domains, often resulting in improved performance and efficiency and in cost savings. However, they also hold potential to enable considerable benefits in sensitive settings, for example in law enforcement or social and health services. Processes for regulatory compliance and certification of compliance already exist in important areas, such as energy, construction, or trade. It is recommended to extend them to also encompass properties specific to big data applications.

Integrate big data in public and private organisations

4

Increase the exploitation of big data technologies in the health sector

Health is a prime example of a sector where the potential of big data analytics is recognised widely by stakeholders but remains far from being fulfilled. A stronger focus on data-based management and decision making will transform current practices in the health sector, potentially improving transparency, quality, safety, efficiency, and coordination of healthcare and enhancing health-related competences of patients. This potential must not be left unused. The legal and ethical aspects need to be addressed for big data analytics to be used more widely in the health sector.

The collection and use of high-quality health data is a key enabler of evidence-based and personalised medicine. It improves diagnoses, allow early detection of high-risk patients, enables the discovery of interactions between diseases or drugs and the identification of risk factors of disease progression, and enhances patient compliance with treatments. However, to enable the use of big data analytics in healthcare, the legal, ethical, administrative, financial or IT barriers must be overcome. The Swiss health system is relatively decentralised and largely organised at the local level, which creates challenges for legislation and administration seeking to enable nationwide data collection and use. From the perspective of data science, this organisation can result in fragmented, local IT solutions with reduced security and quality. The access to data in the right place at the right time and the interfaces among the actors must be improved considerably.

Switzerland has established electronic patient dossiers (EPDs) and quality measures in the health sector that demonstrate the possibility for federal initiatives in this domain. In its current form, the EPD amounts to an unstructured collection of often scanned documents that lacks structured summaries and indices, harmonisation of data formats, semantic interoperability and standardised terminology. Furthermore, it lacks compatibility with automated data analysis requirements and standardisation guidelines. Still, the EPD has enormous potential if an anonymisation layer is introduced to enable collective patient data analyses. There are privacy-preserving technologies that can help in this regard, such as data enclaves, where data is processed without being directly accessible from the outside, or federated analytics that enable local processing of distributed data to avoid compromising security. However, such technologies need further development and require committees with the relevant ethical, legal, scientific, and statistical knowledge to supervise use of the data.

5
Strengthen policy making and evaluation with big data

The collection of data and the increased availability of advanced data analytics combine to offer a powerful foundation for strengthening the factual foundations for policy making. This makes it possible to increasingly quantify social and economic problems and to evaluate the effectiveness of policies and regulations. This potential should be realised in a way that is both responsible and beneficial.

There is enormous potential for the deployment of big data technologies in many sectors of industry and business as well as in public administration. Specifically, deployment of big data technologies coupled with data collection can be applied to public policy making and the monitoring of policy effectiveness, in sectors such as health, energy, finance, transport, spatial planning or sports. Among other beneficial effects, data collection and evaluation enable effectiveness comparisons with policies in other countries as well as identifying improved policies and best practices. This necessitates careful consideration of privacy and security aspects and their proper handling.

In return, data used for the design and evaluation of policies should be publicly available, which again calls for privacy and anonymisation considerations. Altogether, the relevant offices and entities must be strengthened and expanded to handle the increasing workload and necessary communication. Procedures and mechanisms for implementing this cooperation need to be established and developed.

6
Promote shared data collection, application benchmarks and open-source software

The availability of a variety of freely accessible infrastructure will accelerate value creation from data. To enable more open data, adoption of refined data publication policies is recommended. Likewise, better support for the creation of benchmarks and use cases for applications in different domain sciences is warranted. Open-source software represents an attractive alternative to commercial software with expensive licences. To enable additional open-source functionality and capabilities, including, e.g., next-generation computing infrastructure and machine learning toolkits, additional funding for the development of open-source software is recommended.

The development of big data applications is accelerated by the availability of relevant data. However, the collection of data is generally associated with substantial costs. To guarantee continued collection of valuable data by researchers, it is important to ensure a good cost-benefit ratio. Thus, data collection should be incentivised when it comes with substantial costs, for example by allowing delayed publication to enable initial value creation by data collectors before having to make the data available (akin to the role of patenting as a means of encouraging investment in making inventions), by ensuring that funding agencies cover the costs of complying with the FAIR principles for scientific data management[18] that may well extend beyond the end of a project, and by requiring data publication as a prerequisite for publishing associated research results. A nuanced approach to data sharing is needed that recognises that not all data is equally valuable so that resources are spent wisely.

Applications development can also be accelerated by having benchmarks that encompass anonymised datasets and use cases representing common scenarios in the targeted application domains. Such benchmarks can serve as references for application development and testing and for improving the accuracy and predictive capabilities of algorithms. They can enable more effective large-scale deployment and validation of big data applications.

The increased availability of open-source software can accelerate value creation from data. Thus, the creation of incentives for the sharing of tools as open-source software is recommended. For example, impact on society through the development of open-source software with wide-scale adoption should be rated on par with citation counts in the evaluation of scientific careers. This would not only promote an important public service, but might also attract international talent to Switzerland and become an overall important element in the digitalisation of Switzerland.

[18] The FAIR principles are a set of requirements to make research data and other digital objects findable, accessible, interoperable and reusable. See https://www.go-fair.org/fair-principles and Swiss National Open Research Data Strategy, Swissuniversities (2021)

Update and create adequate regulation

7
Pursue more proactive regulation of big data

While big data technologies are being deployed at a rapid pace, regulation is in its infancy and lags well behind the technological development. The lack of regulation can have adverse effects, including on democracy, on the mental health of youth, on competition as well as on innovation, for instance because of unfair advantages and reduced competition. As regulation plays a key role in avoiding such adverse effects and holds the potential to enable improved big data value creation, it is recommended that across-the-board efforts be made to accelerate the regulatory processes.

Applications of big data have a profound and broad impact on society. Regulation can accelerate responsible value creation while limiting adverse effects. A more proactive approach to regulation therefore promotes responsible value creation, facilitates competition and innovation, and serves democracy better.

The big data divide – the asymmetric relationship between those who collect, store, and analyse big data and those subjected to data collection – is an inevitable consequence of a society that values freedom and diversity. Instead of trying to eliminate it, it is recommended that legislation identifies realistic harms that could result from the big data divide and develops legal safeguards for those who are disadvantaged.

Successful big data applications call for trust and acceptance. When putting in place frameworks under which data can be collected, analysed and used, one should not only insist on the creation of (self-regulatory) standards that balance the interests of companies and customers, but should also empower customers to make informed decisions.

Overall, it is of high importance to develop legal safeguards to compensate harm caused by the big data divide by setting standards for data collection, sharing, and analysis, enabling the protection of groups rendered vulnerable due to the deployment of big data technologies.

8
Advance data privacy and digital sovereignty in big data applications

Deploying big data-based applications incurs risks for the privacy and related rights of individuals. Even if basic legal frameworks are available in Switzerland (new Data Protection Act) and in the EU (General Data Protection Regulation GDPR), compliance with the applicable rules is often challenging. It is recommended to raise the awareness of privacy issues and data protection rules among data scientists and engineers, data owners, and data protection officers, to elaborate comprehensive data privacy standards and to pay increased attention to the security of digital infrastructures.

Regulation based on the sovereignty of states over a finite physical space falls short when it comes to big data. Instead, international coordination and cooperation is needed to safeguard the security of digital infrastructures and the privacy and other data-related rights of citizens. Nonetheless, efforts can and should also be made at a national level to ensure data privacy. National and cantonal policy makers and administrations as well as universities and scientists are required to strengthen and complement the national legal framework.

A national data protection and data-related rights agenda includes numerous actors and topics. Therefore, it is important to encourage the creation of strong ties among the many stakeholders from the wide range of disciplines involved in the creation of big data applications. In addition, a methodology should be developed for establishing best practices for gathering and anonymising data, providing secure storage data, and ensuring privacy-preserving value extraction. For example, the development of data enclaves for highly sensitive data appears to be a valuable option to ensure privacy. Furthermore, the existing concept of informed consent should be complemented by specific protection mechanisms.

Various privacy-preserving techniques are being researched, but their real-world deployment will take time. Some of them could contribute to a national data privacy and data-related rights strategy, such as for example the appointment of data trustees protecting personal data of individuals, or the implementation of fairness criteria related to big data analytics to avoid discrimination. To facilitate the daily handling of privacy-related issues, further data protection guidelines should be developed. An expert competence centre such as, or within, the National Cyber Security Centre, offering a public service for addressing legal questions around privacy issues in big data deployment could be also established.

9
Increase transnational harmonisation of regulations

Data often flows across borders, and data access from abroad and international deployment of big data-based services are prevalent. Therefore, a purely national perspective on the application and regulation of big data is insufficient. Rather, it is necessary to observe and engage internationally. Due to the numerous international organisations with their headquarters in Switzerland, Switzerland is in a unique position to support harmonisation activities of transnationally oriented institutions. Switzerland has the opportunity to demonstrate its commitment and expertise in international organisations as well as in national legislation.

The globalisation of data flows and the increased deployment of big data-based applications make it necessary to establish harmonised cross-border regulatory frameworks that cover international trade. While negotiations in the World Trade Organization (WTO) are still ongoing, bilateral and regional (preferential) trade agreements increasingly regulate trade in digital trade and services, as well as data flows. The new rules often encompass aspects of data protection, cybersecurity, and commercial confidentiality.

While Switzerland engages actively in negotiations, further support is recommended. Similarly, it is highly recommended that Switzerland provides its input in the ongoing development of the OECD Guidelines on Responsible Business Conduct (RBC). Finally, Switzerland played an important role as promoter of the UN Internet Governance Forum in Geneva; in view of the increased tensions in the digital world, it is recommended that Switzerland makes efforts to help avoid fragmentation in the regulation of the data-driven economy.

Appendix: The National Research Programme "Big Data" (NRP 75)

www.nrp75.ch

Key facts

Timeline

2014
Proposal of a National Research Programme (NRP) on big data

2015
Mandate by the Federal Council to the Swiss National Science Foundation to conduct NRP 75

2015
Call for research projects and selection

2017–2021
Research work

2022
Synthesis work and dissemination of the results

2023
Publication of the Programme Résumé of NRP 75

Numbers

Budget
CHF 25 million

Projects
34 research projects and 3 cross cutting activities

Organisation

NRP 75 Steering Committee

Professor Christian S. Jensen
Department of Computer Science, Aalborg University (President)

Professor Sihem Amer-Yahia
CNRS, Laboratoire d'Informatique de Grenoble LIG, Université Grenoble Alpes UGA (since 12.07.2016)

Professor Sabrina de Capitani di Vimercati
Computer Science Department, University of Milan

Professor Friedrich Eisenbrand
Institute of Mathematics, EPFL (since 01.01.2021)

Professor Joerg Huelsken
Swiss Institute for Experimental Cancer Research ISREC, EPFL

Professor emeritus Erkki Oja
Department of Computer Science, Aalto University

Professor Reinhard Riedl
Digital Technology Management Institute, Bern University of Applied Sciences

Professor Caroline Sporleder
Institute for Digital Humanities, Georg-August-University of Göttingen (until 31.12.2019)

Professor Rolf H. Weber
Faculty of Law, University of Zurich

Delegate of the Division Programme of the National Research Council for NRP 75

Professor Bert Müller
Biomaterials Science Center, University of Basel (since 01.01.2021)

Professor Friedrich Eisenbrand
Institute of Mathematics, EPFL (until 31.12.2020)

NRP 75 Programme Manager

Boris Buzek
Swiss National Science Foundation, Berne (since 01.11.2022)

Dr Stefan Husi
Swiss National Science Foundation, Berne (from 1.11.2020 to 31.10.2022)

Dr Christian Mottas
Swiss National Science Foundation, Berne (until 31.10.2020)

Representative of the Confederation in NRP 75

Dr Uwe Heck
Federal Chancellery, Digital Transformation and ICT Steering Division (since 01.01.2019)

Willy Müller
Federal IT Steering Unit (until 31.12.2018)

Head of Knowledge Transfer

Beatrice Huber
Swiss Academy of Engineering Sciences (SATW), Zurich (since 01.12.2018)

Dr Béatrice Miller
Swiss Academy of Engineering Sciences (SATW), Zurich (until 30.11.2018)

The 34 research projects

Module 1: Big data Infrastructures

Data streams: monitoring in real time
David Basin, Dmytro Traytel, ETH Zurich
Big data monitoring

Stream analytics: fast processing and privacy-preserving tools
Michael Böhlen, University of Zurich
Privacy preserving, peta-scale stream analytics for domain-experts

Machine learning models: robustness and generalisability
Volkan Cevher, EPFL
Theory and methods for accurate and scalable learning machines

Data centres: efficient performance monitoring
Lydia Yiyu Chen, Delft University of Technology (formerly IBM Research, Zurich)
Dapprox: dependency-ware approximate analytics and processing platforms

Loosely structured data: new tools for integration
Philippe Cudré-Mauroux, University of Fribourg
Tighten-it-all: big data integration for loosely-structured data

Language models: new methods for conversational agents
Thomas Hofmann, ETH Zurich
Conversational agent for interactive access to information

City digital twins: 3D models from a scanning car
Frédéric Kaplan, EPFL
ScanVan – a distributed 3d digitalization platform for cities

Coresets: big data with less data
Andreas Krause, ETH Zurich
Scaling up by scaling down: big ML via small coresets

Scala programming language: enabling big data analytics
Martin Odersky, EPFL
Programming language abstractions for big data

In-network computing: solutions for graph analytics
Robert Soulé, Università della Svizzera italiana
Exploratory visual analytics for interaction graphs

Fast prediction algorithms
Marco Zaffalon, Istituto Dalle Molle di studi sull'Intelligenza Artificiale USI-SUPSI
State space Gaussian processes for big data analytics

Graph analytics and mining
Willy Zwaenepoel University of Sydney (formerly EPFL)
Building flexible large-graph processing systems on commodity hardware

Module 2: Societal and regulatory challenges

Trade agreements: impacts on national law
Mira Burri, University of Lucerne
The governance of big data in trade agreements: design, diffusion and implications

Big data in insurance
Markus Christen, University of Zurich
Between solidarity and personalization – Dealing with ethical and legal big data challenges in the insurance industry

Regulating big data research
Bernice Simone Elger, University of Basel
Ethical and legal regulation of big data research – Towards a sensible and efficient use of electronic health records and social media data

Legal challenges of big data
Sabine Gless, University of Basel
Legal challenges in big data. Allocating benefits. Averting risks

Uncertainty in big data applications: lessons from climate simulations
Reto Knutti, ETH Zurich
Combining theory with big data? The case of uncertainty in prediction of trends in extreme weather and impacts

Big data in practice: sociology, data sciences and journalism
Sophie Mützel, University of Lucerne
Facing big data: methods and skills needed for a 21st century sociology

Big data in health: an ethical framework
Effy Vayena, ETH Zurich
BEHALF – Bigdata-ethics-health framework

Big data in human resources
Antoinette Weibel, University of St.Gallen
Big data or big brother? Big data HR control practices and employee trust

Module 3: Big data applications

Optimising transport management: anonymous individual mobility traces
Kay W. Axhausen, ETH Zurich
Big data transport models: the example of road pricing

Pig Data: analytics for the Swiss swine industry
John Berezowski, University of Bern
Pig data: health analytics for the Swiss swine industry

Flood detection: automatic geotagging of crowdsourced videos
Susanne Bleisch, FHNW
EVAC – Employing video analytics for crisis management

Computational chemistry: discovering new molecules
Helmut Harbrecht, University of Basel
Big data for computational chemistry: unified machine learning and sparse grid combination technique for quantum based molecular design

Intensive care units: an automated alert system
Emanuela Keller, University Hospital Zurich
ICU-cockpit: IT platform for multimodal patient monitoring and therapy support in intensive care and emergency medicine

Evidence-based policy: uncovering causality from data
Michael Lechner, University of St. Gallen
Causal analysis with big data

Mapping global innovation: analysing patents
Alessandro Lomi, Università della Svizzera italiana
The global structure of knowledge networks: data, models and empirical results

Big genetic data: powerful indexing
Gunnar Rätsch, ETH Zurich
Scalable genome graph data structures for metagenomics and genome annotation

Back pain management: a personalized smartphone-based solution
Robert Riener, ETH Zurich, and Walter Karlen, Ulm University (formerly ETH Zurich)
Personalized management of low back pain with mHealth: big data opportunities, challenges and solutions

Soil erosion: quantification by aerial photography in Switzerland
Volker Roth, University of Basel
WeObserve: integrating citizen observers and high throughput sensing devices for big data collection, integration, and analysis

Genome comparison: faster analysis
Nicolas Salamin, University of Lausanne
Efficient and accurate comparative genomics to make sense of high-volume low-quality data in biology

Renewable energy potential: evaluation for Switzerland
Jean-Louis Scartezzini, EPFL
Hybrid renewable energy potential for the built environment using big data: forecasting and uncertainty estimation

Bioinformatics databases: queries in natural language
Kurt Stockinger, ZHAW
BIO-SODA: enabling complex, semantic queries to bioinformatics databases through intuitive searching over data

Solar eruptions: predicting geomagnetic storms
Svyatoslav Voloshynovskiy, University of Geneva
Machine learning based analytics for big data in astronomy

The 3 cross-cutting activities

Big data: open data and legal strings
Sabine Gless, University of Basel

ELSI Task Force for the National Research Programme “Big Data”
Markus Christen, University of Zurich

Women in big data
Lydia Yiyu Chen, Delft University of Technology (formerly IBM Research)

Publications and teaching material

Ethical, legal and social issues of big data – A comprehensive overview
Eleonora Viganò (Ed.), NRP 75 (2022)

Big data: outil pédagogique pour les cycles secondaires; Big Data: Lehrmittel für die Sekundarstufen NRP 75 and Museum of Communication Bern (2020)

Big data ethics recommendations for the insurance industry, NRP 75 (2019)

Imprint

This Résumé of the National Research Programme “Big Data” (NRP 75) summarises the findings of the NRP 75's 37 projects, integrating them into an overview of the opportunities and challenges of big data. The authors synthesised these findings, bringing their scientific expert knowledge and experience. The text was consolidated and edited by a science journalist. The conclusions presented at the end of the document are the result of a collective multi-stage process: they were drafted, discussed and consolidated by members of the Steering Committee, and represent the approved consensus of the Steering Committee.

The Résumé is to be considered as a scientific contribution to the process of opinion formation, to political and specialist debate as well as to the planning of strategies and measures for the political and social transformations with and through big data. The text is the collective responsibility of the Steering Committee. Their assessments and conclusions do not necessarily reflect those of the research teams or the Swiss National Science Foundation. Further information on all of the NRP 75 research projects named in the Résumé can be found on the website www.nrp75.ch.

Authors

Professor Sihem Amer-Yahia, Université Grenoble Alpes UGA

Professor Joerg Huelsken, EPFL

Professor Christian S. Jensen, Aalborg University

Professor emeritus Erkki Oja, Aalto University

Professor Reinhard Riedl, Berne University of Applied Sciences

Professor Rolf H. Weber, University of Zurich

Editor

Dr Daniel Saraga, Saraga Communications, Basel

Coordination

Boris Buzek, Swiss National Science Foundation, Berne

Dr Stefan Husi, Swiss National Science Foundation, Berne

Beatrice Huber, Head of Knowledge Transfer NRP 75

Production

Sub-editing and proofreading:
Edwin Cartlidge, Rome

Layout and illustrations:
Gabriel Alber, Alber Visuelle Kommunikation, Zurich

Printing:
Druckerei Herzog AG, Langendorf (SO)

Publisher

Steering Committee of the National Research Programme “Big Data” (NRP 75)

Suggested citation:

Steering Committee of NRP 75 (2023): “Big Data: Applications, technologies, and societal aspects. Résumé of the National Research Programme ‘Big Data’ (NRP 75)”, Swiss National Science Foundation, Berne.

National Research Programmes (NRP) are a funding instrument of the Swiss National Science Foundation (SNSF). Their aim is to enable research contributing to the solution of contemporary problems of national importance. The National Research Programme “Big Data” (NRP 75) ran from 2015 to 2022.

ISBN: 978-3-907087-61-9